NS1.20: W 89/v.2

World Ionosphere/ Thermosphere Study

WITS HANDBOOK

Volume 2

Edited by

C. H. Liu

a.f.

International Council of Scientific Unions (ICSU)

Scientific Committee on Solar-Terrestrial Physics (SCOSTEP)
J. G. Roederer, President
W. I. Axford, Vice President
C. H. Liu, Scientific Secretary

WITS ORGANIZATION

World Ionosphere/Thermosphere Study (WITS) Steering Committee

K. D. Cole and C. H. Liu, Cochairmen
H. C. Carlson
G. S. Ivanov-Kholodny
S. Kato
A. P. Mitra
J. Röttger

WITS PROJECTS

Global Irregularity Study (GIS)
B. G. Fejer, Coordinator

Global Ionosphere-Thermosphere Coupling and Dynamics (GITCAD)
V. B. Wickwar, Coordinator

Instabilities and Nonlinear Effects in the Ionosphere/Thermosphere System (INEITS)
V. Trakhtengerts and E. Mishin, Coordinators

Ionospheric Perturbations and Geomagnetic Pulsations
K. D. Cole, Coordinator

Lower Thermosphere Coupling Study (LTCS)
J. M. Forbes, Coordinator

Solar Electromagnetic Radiation Flux Study (SERFS)
R. F. Donnelly, Coordinator

SUNDIAL
E. P. Szuszczewicz, Coordinator

Real-Time Support of WITS Campaigns
J. A. Joselyn, Coordinator

World Acoustic Gravity Waves (WAGS)
R. D. Hunsucker and P. Argo, Coordinators

World Ionosphere/ Thermosphere Study

WITS HANDBOOK

Volume 2

Edited by

C. H. Liu

December 1989

Published for the ICSU Scientific Committee on Solar-Terrestrial Physics (SCOSTEP)
with financial assistance from Unesco subvention through a 1989 grant from ICSU

Copies available from the SCOSTEP Secretariat,
University of Illinois, 1406 W. Green St., Urbana, Illinois 61801

II

FOREWORD

The current resurgence of interest in the study of the ionosphere-thermosphere system is, to a significant degree, due to the development and improvement of experimental and analysis techniques. These include space-borne as well as ground-based systems, carrying out either *in situ* or remote sensing measurements, actively or passively, using a significant fraction of the electromagnetic spectrum. Innovations are found not only in hardware, but also in methods of analysis and simulations. During the period of WITS, it was recognized that a Handbook dedicated to state-of-the-art technical information would be most useful to help workers in the field keep abreast of new developments. To this end, recognized experts have been invited to contribute chapters on their specialities, covering such aspects as instrumentation, technical principles, data analysis, and capabilities. Due to space limitations it has not been possible to cover all topics exhaustively. It is hoped that future volumes dedicated to techniques will be published.

I would like to thank all the contributors for their tremendous effort and cooperation. Special thanks are also due to Mrs. Belva Edwards for her editorial help.

C. H. Liu

Table of Contents

Chapter Page

1. Use of ground-based magnetometers to infer the dynamics of magnetosphere-ionosphere coupling and the solar-terrestrial interaction, *G. Rostoker*........1

2. "In situ" measurement techniques for ionospheric-thermospheric investigations, *E. P. Szuszczewicz*........44

3. Prospects of topside sounding, *S. A. Pulinets*........99

4. Scintillation technique for probing ionospheric irregularities, *Sa. Basu and Su. Basu*........128

5. Modern total electron content measurement techniques, *J. A. Klobuchar*........137

6. Rayleigh and sodium lidar techniques for measuring middle atmosphere density, temperature and wind perturbations and their spectra, *C. S. Gardner, D. C. Senft, T. J. Beatty, R. E. Bills, and C. A. Hostetler*........148

7. Upper atmospheric wind and temperature measurements using imaging Fabry-Perot interferometers, *D Rees , I. McWhirter, A. Aruliah, and S. Batten*........188

8. Spectrometric measurement techniques of weak emissions in twilight airglow, *C. A. Tepley*........224

9. Optical and radar techniques applied to chemical research "active experiments" in the ionosphere/thermosphere system, *M. Mendillo, J. Baumgardner, and P.J. Sultan*........250

10. Modification of the ionosphere as a method of its diagnostics, *L. M. Erukhimov and N. A. Mityakov*........267

11. Coherent radar techniques, *E. Nielsen*........285

12. Remote sensing of the high latitude ionosphere with coherent scatter HF radar: A review of the technique and recent work, *J. M. Ruohoniemi, R. A. Greenwald, and K. B. Baker*........319

13. The digisonde 256 ionospheric sounder, *B. W. Reinisch, K. Bibl, D. F. Kitrosser, G. S. Sales, J. S. Tang, Z.-M. Zhang, T. W. Bullett and J. A. Ralls*........350

14. The HF Doppler technique for monitoring transient ionospheric disturbances, *T. B. Jones*........383

15. Theoretical and simulation methods applied to high latitude, F region turbulence, *J. D. Huba*........399

16. The derivation of meridional neutral winds in the thermosphere from F2-layer height, *K. L. Miller, P. G. Richards, and D. G. Torr*........439

Chapter 1

USE OF GROUND BASED MAGNETOMETERS TO INFER THE DYNAMICS OF MAGNETOSPHERE-IONOSPHERE COUPLING AND THE SOLAR-TERRESTRIAL INTERACTION

Gordon Rostoker

Institute of Earth and Planetary Physics and Department of Physics
University of Alberta, Edmonton, Alberta, Canada T6G 2J1

ABSTRACT

For many years, the level of magnetospheric and ionospheric activity has been quantified by the use of ground based measurements of the perturbation magnetic field at selected positions on the Earth's surface. In more recent times, our understanding of how to interpret the magnetic perturbation pattern has allowed researchers to move from the use of crude indices of activity to detailed inferences from arrays of magnetometer data and from individual magnetograms themselves. In this paper, I shall describe how magnetometer data were used in the early days of space research to provide insight into magnetospheric dynamics and solar-wind magnetosphere coupling. I shall then outline modern techniques for conversion of raw magnetometer data into information about ionospheric currents and the field-aligned currents which couple the ionosphere to the distant magnetosphere. I shall also provide schemes by which researchers can use individual magnetograms to infer the position of the auroral oval and the strength of the electric currents flowing along the oval.

INTRODUCTION

At the present time, there are about 100 magnetometers operating continually at observatories distributed irregularly over the Earth's surface. Most of these observatories are sited for reasons other than scientific logic so that one is not dealing with an organized array of instrumentation. Nonetheless, these magnetometers together with some small organized arrays operating in an expeditionary fashion have provided the space research community with an invaluable tool for studying the solar-terrestrial interaction over the past few decades. In the early part of the twentieth century, most research involving ground magnetometer data involved statistical studies of data from single stations. Starting with the pioneering work of Silsbee and Vestine [1942] and Harang [1946], efforts were made to take advantage of accidental arrangements of instruments (eg. magnetometers lying along a line of constant magnetic latitude or longitude) to infer information about localized current distributions in the ionosphere.

The beginning of modern space research is marked by the publication of two foundation stone papers at the start of the 1960's. Around that time, Axford and Hines [1961] and Dungey [1961] presented physically viable pictures of how energy from the solar wind could penetrate the inner magnetosphere and eventually be deposited in the upper atmosphere. The scheme of energy transport within the magnetosphere developed by Axford and Hines was based on the character of the DS equivalent current system, which, in turn, was derived using ground magnetometer data. In fact, countless papers on ionospheric and upper atmosphere research in the early years of space science centered on the estimation of the equivalent current systems for various types of geomagnetic perturbations. One of the first things we shall do in this paper is to describe equivalent current systems and the means by which they are derived.

By the late 1960's some researchers began to set up specific arrays of magnetometers in order to study the dynamic behaviour of the auroral electrojets. Following the research of Bonnevier et al. [1970] using a fortuitous alignment of stations along a line of constant longitude across the average position of the auroral oval, Rostoker et al. and Akasofu et al. set up and operated meridian lines of magnetometers in Canada and Alaska respectively. Some two dimensional arrays were also operated on an expeditionary basis during the 1970's which were used for the study of upper atmospheric current systems (e.g. Bannister and Gough, 1977). While meridian lines of magnetometers were useful in delineating the latitudinal structure of the auroral electrojets, it was ultimately necessary to establish east-west aligned magnetometer lines to study the azimuthal structure in the electrojets and in the associated auroral luminosity. Such arrays were constructed by Rostoker et al. and by Baumjohann et al. in Canada and Scandinavia respectively and have been particularly useful in the study of ULF magnetic pulsations and substorm expansive phase activity. In the second part of this paper, I shall outline the data presentation modes for meridian line and east-west line data and suggest inferences which can be made from such data sets regarding ionospheric and field-aligned currents.

During the 1980's, large computers began to play an increasingly important role in the process whereby researchers convert knowledge of ground based magnetometer data into information about the real currents flowing in the magnetosphere-ionosphere system. Taking advantage of this evolving computational tool, Richmond et al. in the U.S.A. and Mishin et al. in the U.S.S.R. began to develop algorithms for evaluating the instantaneous real current distributions using as input data the ground magnetometer data and a model ionospheric conductivity distribution. In the final part of this paper, I shall outline the technique for such current system estimations following the approach of Kamide et al. [1981], which is known as the KRM method.

MEASUREMENTS OF THE PERTURBATION MAGNETIC FIELD AT THE EARTH'S SURFACE

It would be ideal if all magnetometers recording at the Earth's surface used the same coordinate system in which to represent their data. Unfortunately, such is not the case with the data being recorded in two primary coordinate systems and subsequently being portrayed in the literature in several other coordinate systems after appropriate coordinate transformations. The two primary systems for the recording of data are:

(1) Geographic where the X component is defined as positive northward, the Y component is defined as positive eastward and the Z component is defined as positive downward.

(2) Local magnetic where the H component is defined as positive northward, the D component is defined as positive eastward and the Z component is defined as positive downwards.

While the X component lies in the geographic meridional plane, the H component lies in the local magnetic meridian plane. Figure 1 shows the geometrical arrangement of the relevant coordinate axes for the geographic and local magnetic coordinate directions. In addition, it should be noted that researchers often portray their perturbation vectors in the centered dipole coordinate system (X_m, Y_m, and Z) where X_m is positive pointing toward geomagnetic north, Y_m is positive towards geomagnetic east and Z is positive downwards. The horizontal axis directions are also shown in Figure 1 and the coordinate transformation equations from local magnetic to geographic and local magnetic to centered dipole are respectively

$$\begin{aligned} \Delta X &= \Delta H \cos D - (H_o \Delta D) \sin D \\ \Delta Y &= \Delta H \sin D - (H_o \Delta D) \cos D \end{aligned} \qquad (1)$$

$$\begin{aligned} \Delta X_m &= \Delta H \cos (D - \Psi) - H_o \Delta D \sin (D-\Psi) \\ \Delta Y_m &= \Delta H \sin (D - \Psi) + H_o \Delta D \cos (D-\Psi) \end{aligned} \qquad (2)$$

where H_o is the total horizontal field strength, D is the magnetic declination measured in degrees and Ψ is the great circle distance from the geomagnetic to the geographic pole measured in degrees. Other more elegant coordinate systems such as corrected geomagnetic coordinates (taking into account the quadrupole components of the main field) are also used from time to time (viz. Hakura, 1965); however these are more difficult to deal with since the systems may not be orthogonal. The centered dipole system appears to be the most effective orthogonal coordinate system in which to order ground magnetometer data to this point in time.

In general, one categorizes magnetometer locations as either high latitude, auroral zone or low latitude. Implicit in these categorizations is the assumption that:

(1) High latitude means that the observing station is poleward of the high latitude edge of the auroral oval. (One cannot assign a specific lower latitude to the "high latitude" region, because the auroral oval grows and shrinks with changing levels of magnetospheric activity.)

(2) Auroral zone means that the observing station lies inside the auroral oval (i.e., ionospheric electrojet current flows over the observing station).

(3) Low latitude pertains to the region extending from the equatorward border of the auroral oval to the Earth's magnetic equator.

Although the term "middle latitude" also is found in the literature, there is no physical basis for breaking up the region between the equatorward edge of the auroral oval and the geomagnetic equator into two distinctive regions. There is some reason to believe that some interesting physical processes take place at the plasmapause (which lies equatorward of the auroral oval); however the term "middle latitude" does not apply to a region one of whose boundaries is the plasmapause.

Figure 2 shows typical magnetograms from stations located in the three regions described above. In this case, the perturbations are measured in the geographic coordinate system but in central Canada (where the observatories are located), the local magnetic north and geographic north directions are almost the same. These data were taken digitally and have been subsequently put into the analogue form shown here. However, there still are many stations in the global network which record on strip chart. The different recording techniques used at the various observatories around the world makes quantitative analysis of the magnetic perturbation data more difficult than one would like. Inspection of Figure 2 reveals that the magnetic perturbation pattern for the same

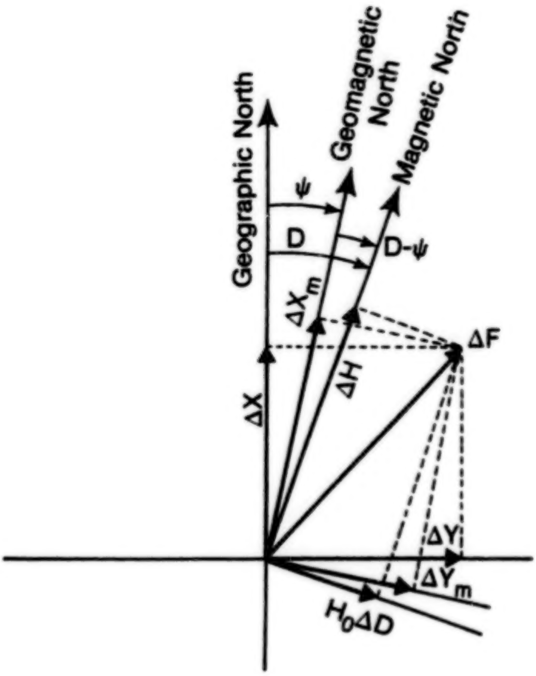


Figure 1. Relationships among the various coordinate systems used to display ground based geomagnetic data. The coordinate transformations between systems are shown in Eq. (1).

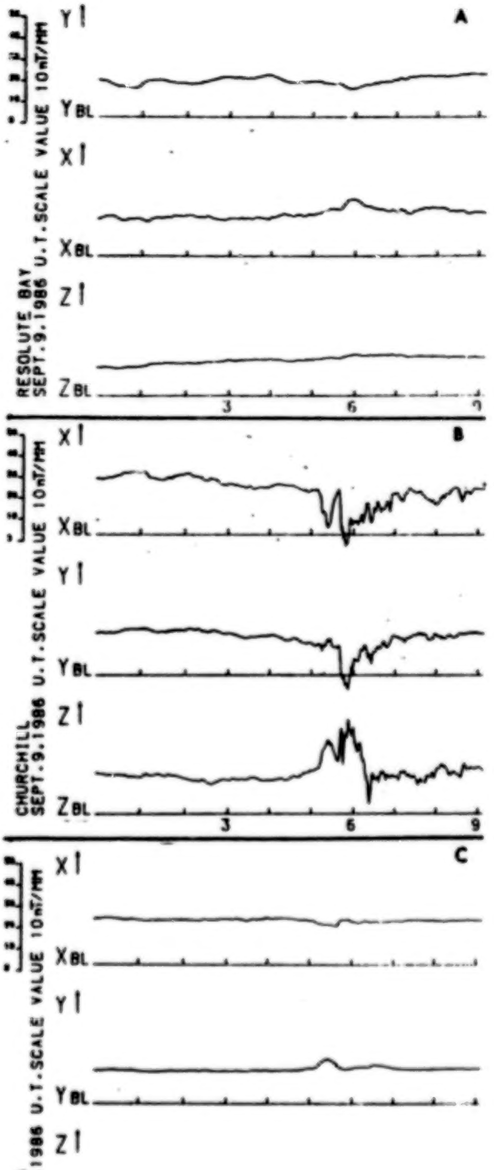


Figure 2. Three typical magnetograms displaying a substorm event as viewed at high latitude (A), auroral oval (B) and subauroral (C) stations. Magnetospheric substorm disturbances are seen between 0500 and 0700 UT. Stations well outside the oval see predominantly the effects of field-aligned currents.

period of time is quite different in the three different zones alluded to above. This demonstrates unambiguously that the electric current systems being monitored can be spatially quite localized.

One can immediately appreciate that a full description of an individual event would require the presentation of an unreasonable amount of magnetometer data if these data had to be expressed in the form shown in Figure 2. For this reason, researchers developed methods of portraying the essential information in a more compact form. The first of these modes of presentation to gain favour was the equivalent current system. Figure 3 shows a typical equivalent current system for relatively disturbed times. In the construction of such a figure, it is assumed that the magnetic perturbations everywhere are due solely to ionospheric currents. The strength of the current over each site is estimated by assuming that the horizontal magnetic perturbation vector is due to an infinite sheet overhead current with often no correction for Earth induction being considered. The current vectors shown on the figure have a magnitude proportional to the inferred current density and the current vectors are oriented normal to the magnetic perturbation vectors. The equivalent current contours are sketched in a fashion so that the strength of the equivalent current is inversely proportional to the distance between the current lines. This mode of presentation is compact, however it contains no information regarding the vertical component of the perturbation magnetic field. As we shall see, the Z-component of the perturbation magnetic field can be extremely useful in inferring the properties of the real currents flowing in the magnetosphere-ionosphere system.

Another way in which magnetospheric activity levels have been quantified using magnetometer data is through the development of geomagnetic indices. The primary ones used in the past are the K_p, Dst and AE indices. (The reader is referred to Rostoker [1972] for a detailed description of the derivation of these indices.) At the present time, only the AE index is used extensively so that I shall only summarize the origins of that index. The process of creating the index involves the use of the magnetograms from 12 stations distributed in a relatively uniform fashion around the world at average auroral zone latitudes. The traces of the north-south component are superposed on one another and at each minute, the maximum positive value and the maximum negative value of the 12 traces are chosen. The time series of maximum positive values is called AU while the time series of maximum negative values is called AL. The sum of the absolute values of AU and AL at each minute has been termed AE. A sample daily record of the AU, AL and AE indices is shown in Figure 4. It is important to realize that the time series of the AU, AL and AE indices have no physical meaning as they do not reflect the intensity variation of any particular magnetosphere-ionosphere current system. The AU index is a crude estimator of the lower limit of the strength of eastward flowing ionospheric current (which creates a positive deflection in the north-south component of the perturbation magnetic field below the ionosphere) while the AL index reflects the strength of the westward flowing ionospheric current (creating a negative deflection in the north-south component of the perturbation magnetic field). The AE index crudely quantifies activity level but yields no information about what current systems may cause that activity.

There are several weaknesses in the formulation of the AU, AL and AE indices which makes their use risky in any way other than establishing a lower limit to the magnetospheric activity level. The problems are:

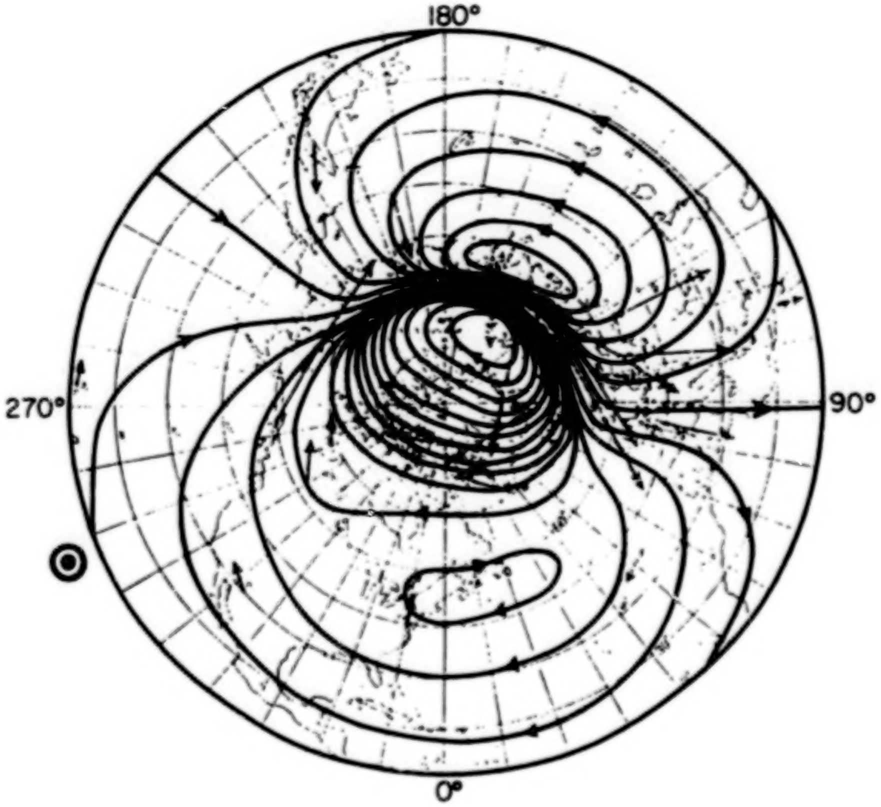


Figure 3. Equivalent current system for a polar magnetic substorm (Fukushima, 1953). The distance between current lines is approximately inversely proportional to the strength of the equivalent overhead current.

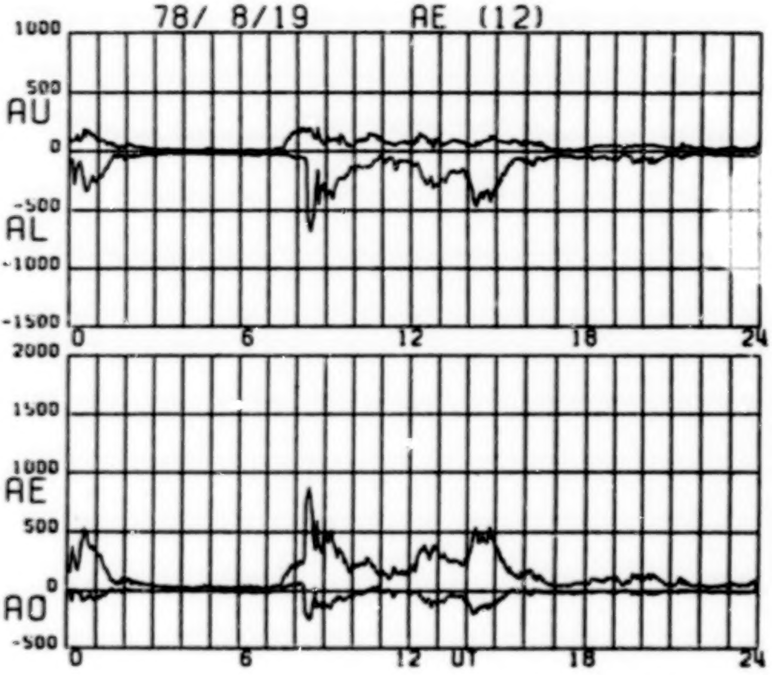


Figure 4. Typical records of AE, AU and AL over the period of one day during which magnetospheric substorm activity was in evidence.

1. There is a high probability that no AE station is located at the locale of the true maximum magnetic perturbation at the earth's surface. This is because, during quiet times, the contracted auroral oval typically lies poleward of the AE stations while, during active time the currents flowing in the oval may lie equatorward of the AE stations. Figure 5 shows values of the AE index for one day where the normal AE (created from the data of 12 stations) could be compared with an AE index created using a network of 70 stations. It is evident that the magnetic activity level can often be much larger than the normal AE values might suggest (which is particularly evident from the two AE traces in the Universal Time interval 1700 - 2300 on March 18, 1978).

2. It is usual for the researcher to associate the geomagnetic activity index with the strength of convection (and hence the strength of the convection electric field) in the magnetosphere. However, the magnetic perturbations reflect the convolving of electric field and conductivity at each point in the ionosphere. During significant enhancements of the magnetosphere-ionosphere interaction, the conductivity can become markedly enhanced in the midnight sector so that large ionospheric currents can be produced with relatively weak electric fields.

3. The magnetic perturbations of field-aligned currents are superposed on the magnetic fields produced by the ionospheric currents, tending to reduce the estimate of the overhead current being remotely sensed by the AE stations. Figure 6 shows a cartoon of a three-dimensional current system believed to flow during major auroral outbursts in the midnight sector. It is clear that the magnitude of the north-south component of the perturbation magnetic field due to the ionospheric current is reduced due to contributions from the field-aligned currents linking the ionospheric currents to the outer magnetosphere.

4. There is no way in which the changes in magnetic field due to the spatial motion of localized current systems can be separated from changes in the actual magnitudes of the current flow in those systems. For example, a decrease in the AU index may simply reflect the equatorward motion of a latitudinally localized eastward current from directly overhead to some position well distant from the AE station.

For these reasons, any researcher wishing to study the details of some individual event in terms of establishing the nature of the magnetosphere-ionosphere current system ought not to use the indices but rather should study the individual magnetograms from relevant stations.

THE ELECTRIC CURRENT SIGNATURES OF THE COMPONENTS OF MAGNETOSPHERIC SUBSTORM ACTIVITY

The electric current systems which couple the distant magnetosphere to the ionosphere play a vital role in regulating the energy flow within the magnetosphere and the ultimate degradation of high grade bulk flow energy into low grade heat. It is generally believed that solar wind plasma makes its way from the magnetosheath into the boundary layers just inside the magnetopause (cf. Eastman et al., 1985). Within the boundary layers, the convective bulk flow energy of the plasma is extracted through magnetohydrodynamic (MHD) generator processes (cf. Lundin, 1988) and converted to the electromagnetic energy of the three-dimensional current systems which couple the ionosphere and magnetosphere. Ultimately, this electromagnetic energy is converted to heat through

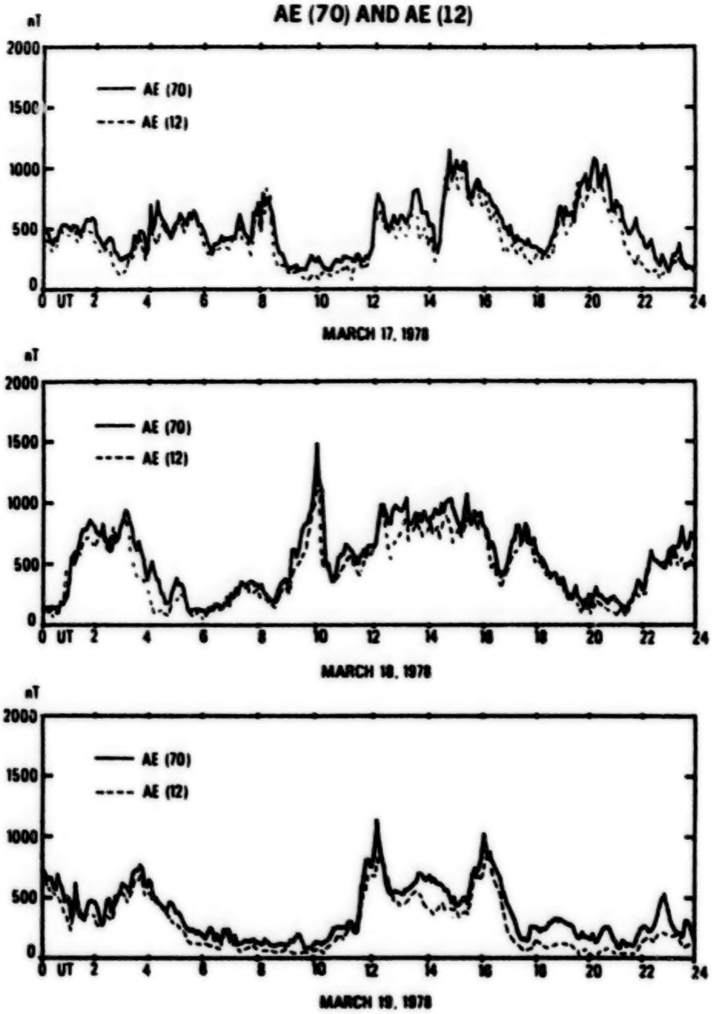


Figure 5. Comparison of AE computed using data from the 12 normal stations with AE computed using data from 70 ground observatories distributed over the Earth's surface (Kamide et al., 1982). The 12 station index clearly underestimates the activity level, particularly in the interval after 1800 UT on March 19, 1978.

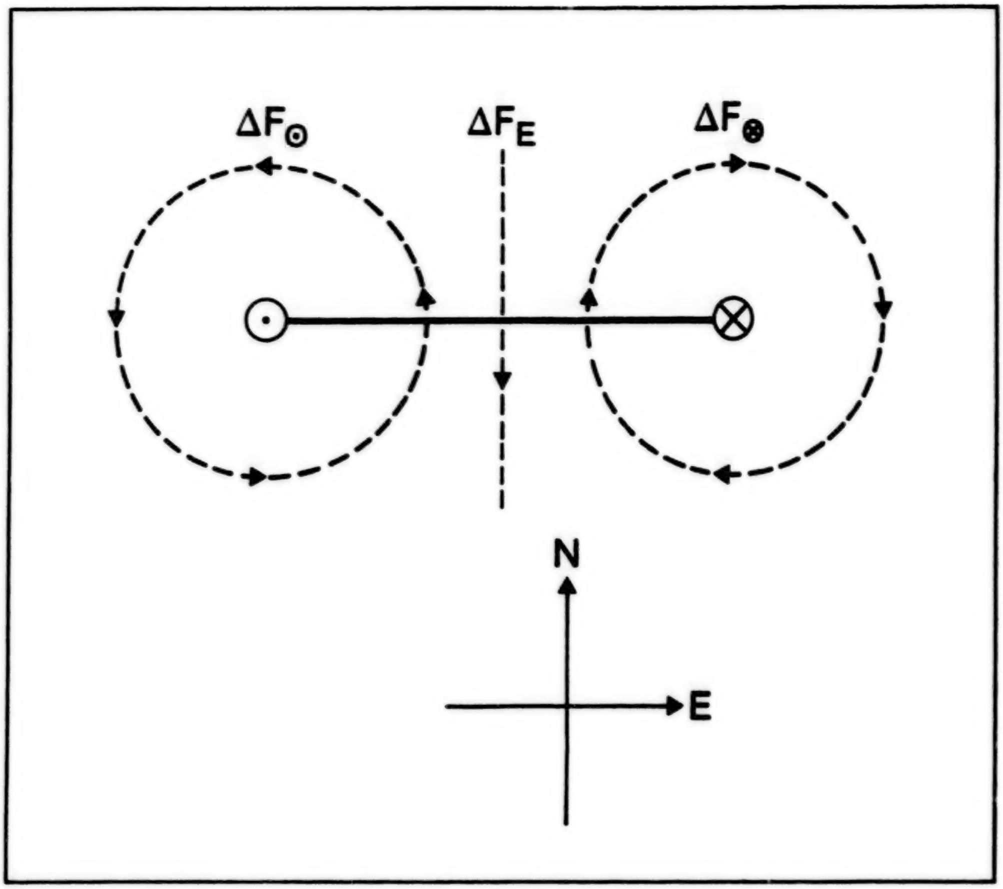


Figure 6. Equivalent three dimensional current system associated with the substorm current wedge observed during expansive phase activity. The magnetic perturbations from the field-aligned and ionospheric currents clearly oppose one another directly under the ionospheric electrojet.

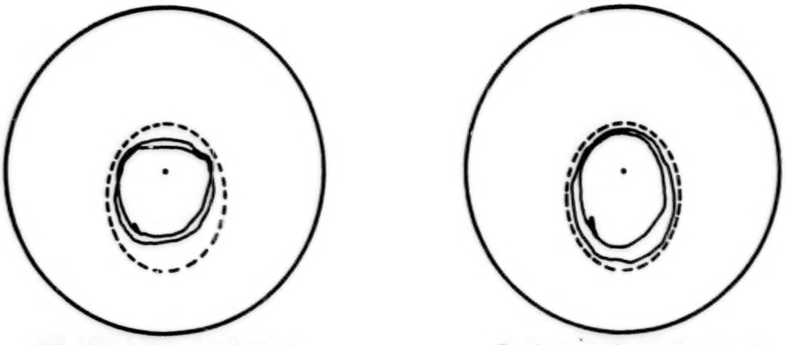
Joule heating associated with Pedersen current flow in the auroral ionosphere. The concurrent heating of plasma in the central plasma sheet through both adiabatic processes (Hines, 1963) and non-adiabatic processes (Goertz and Smith, 1989) leads to pressure gradients. Once again, MHD generator processes lead to both inertial and diamagnetic currents being driven in the central plasma sheet and these can be connected to the auroral ionosphere by Birkeland currents thus allowing energy to be extracted from the magnetospheric plasma (cf. Rostoker and Boström, 1976). In this fashion, the plasma sheet is cooled and the upper atmosphere is heated.

The electric current systems described above fluctuate in response to changes in the interplanetary medium. Perrault and Akasofu [1978] have defined the epsilon parameter which gives a reasonable quantitative measure of the amount of energy which penetrates the magnetosphere in terms of the relevant interplanetary parameters, viz.

$$\varepsilon = \ell_0^2 \, v \, B^2 \sin^4 \frac{\Theta}{2} \tag{3}$$

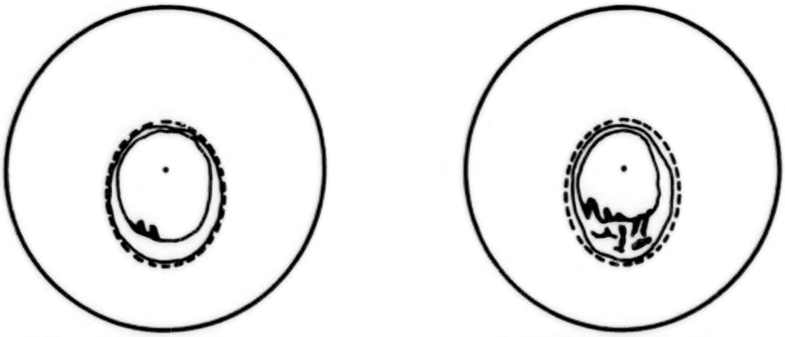
where v is the solar wind speed, B is the magnitude of the interplanetary magnetic field, Θ is the polar angle of the component of the IMF normal to the sun-earth line and measured from the northward geomagnetic axis and $\ell_0 (\sim 7R_E)$ is a constant with the dimensions of distance. While v and B normally do not change by factors of more than two, the angle Θ can change by as much as 180° in a matter of seconds and often does so. Thus the amount of energy entering the magnetosphere is strongly modulated by the direction of the IMF with maximum entrance of energy occurring for $\Theta = 180°$ (purely southward IMF) and minimum entrance of energy occurring for $\Theta = 0°$ (purely northward IMF). It is common to study episodes of enhanced magnetospheric activity, such as those seen in Figure 4. Such episodes have been labeled magnetospheric substorms and the study of substorm magnetic signatures is still a very active research area. For this reason it is useful to spend some time is describing the various aspects of substorm activity.

Figure 7 shows the evolution of auroral activity as the auroral oval experiences an episode of magnetospheric substorm activity. The large scale behaviour of the oval involves an equatorward expansion during the period when the IMF has become more southward and a contraction to its previous location when the IMF returns to its previous orientation. (This is equivalent to ε increasing and subsequently decreasing back to its original value.) During the period of enhanced energy input from the solar wind, the oval can brighten and distorted discrete auroral forms called surges can appear from time to time on its poleward border in the evening sector. At the same time, electric current strength increases along the high conductivity channel marked by the auroral luminosity which defines the oval. The ionospheric current flow is eastward across the dusk meridian and westward across the dawn meridian. These two ionospheric currents along with their associated closure currents in the magnetosphere constitute the directly driven component of substorm activity. At the same time as the directly driven system of currents grows, energy is also stored in the magnetotail with this storage feature of the substorm being called the growth phase. The return of ε to its pre-substorm value can often lead to explosive discrete auroral arc development in the midnight and evening sector of the oval. This often violent auroral activity is accompanied by a poleward motion of the poleward edge of the oval. The beginning of this poleward motion marks the onset of the substorm expansive phase. The expansive phase features many individual arc formations of limited longitudinal extent



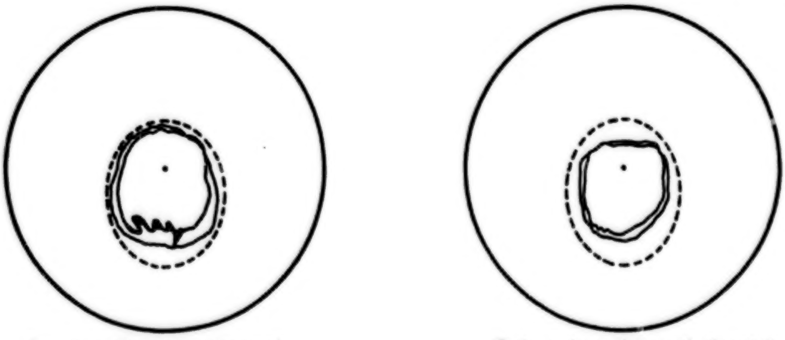
IMF after being northward for some time has just turned southward

Oval expands equatorward under influence of southward IMF



IMF has just turned northward

Poleward edge of oval expands poleward and features multiple surges. Bright structures seen within diffuse evening auroras



Surge action at poleward edge of oval continues as diffuse auroras recede poleward

Return to quiet oval about 1 hr after IMF northward turning

Figure 7. Sequence of plots of the auroral oval during the growth and recovery of a magnetospheric substorm. Auroral surges are found at the poleward edge of the oval, with the size of the polar cap being regulated by the Bz component of the interplanetary magnetic field (Rostoker, 1989).

and featuring a surgelike structure at their western edges. Each of these surges is the site of field-aligned current flowing out of the ionosphere. This field-aligned current is fed predominantly by a westward electrojet coming from regions to the east of the surge form. The combination of the magnetic effects of the upward field-aligned current and the ionospheric currents in its immediate vicinity produces a unique magnetic signature which we shall discuss later in this paper. The immediate cause of the distorted discrete auroral features is thought to be the release of energy from the magnetotail which was stored during the growth phase. The combined storage-release process constitutes the loading-unloading component of substorm activity.

MAGNETIC PERTURBATION PATTERNS OF MAGNETOSPHERE-IONOSPHERE CURRENT SYSTEMS

Before looking at the details of the magnetic signatures of directly driven activity and the unloading of magnetotail energy, it is useful to look at the magnetic perturbations due to some simple current configurations involving ionospheric current flow. Figure 8a shows a latitude profile of an infinite line current flowing westward in the ionosphere at a latitude of 70° N. One finds a negative H-component perturbation directly under the current system which falls off to zero toward lower and higher latitudes. The Z-component reverses polarity across the latitude where the line current flows, with a positive perturbation to the north and a negative perturbation to the south. Had the line current been directed eastward in the ionosphere, the polarities of all the perturbations shown in Figure 8a would have been reversed. Note that, because the line current is of infinite length, there is no D-component perturbation for this system. Figure 8b shows the magnetic perturbation pattern that would have been expected had the ionospheric current flowed from south-west to north-east rather than being purely westward. Now one can see perturbations in both the H-component and D-component, however the peak is located at the same latitude for both components.

We now consider the magnetic perturbations from a three-dimensional current system such as that shown in Figure 9. Profiles for a number of representative meridians are shown in Figure 10 as calculated using the algorithms developed by Kisabeth and Rostoker [1977]. One can immediately see that the extremum values in the Z-component perturbation mark the borders of the westward current flow (which is of finite latitudinal extent) while the polarities of the D-component give information as to whether the observing stations are poleward or equatorward of the center of the westward electrojet and to the east or west of the meridian which bisects the electrojet (which is of finite latitudinal extent). It is also noticeable that the extremum values in the D-component lie just equatorward and poleward of the borders of the ionospheric electrojets, with the magnitude of the perturbation falling off slowly as one moves to lower and higher latitudes respectively. At middle latitudes, the D-component is clearly dominant, the H-component has reversed polarity (as the field-aligned current begins to dominate the ionospheric electrojet current) and one can barely see the edge effect of the electrojet in the weak Z-component perturbation.

MAGNETIC SIGNATURES OF THE DIRECTLY DRIVEN SYSTEM

As we pointed out in the previous section, the directly driven system features an eastward electrojet flowing across the dusk meridian and a westward electrojet flowing across the dawn meridian. The electrojets produce perturbations in the horizontal component of the magnetic field ranging from tens of nT during quiet times to a thousand nT or more during active periods.

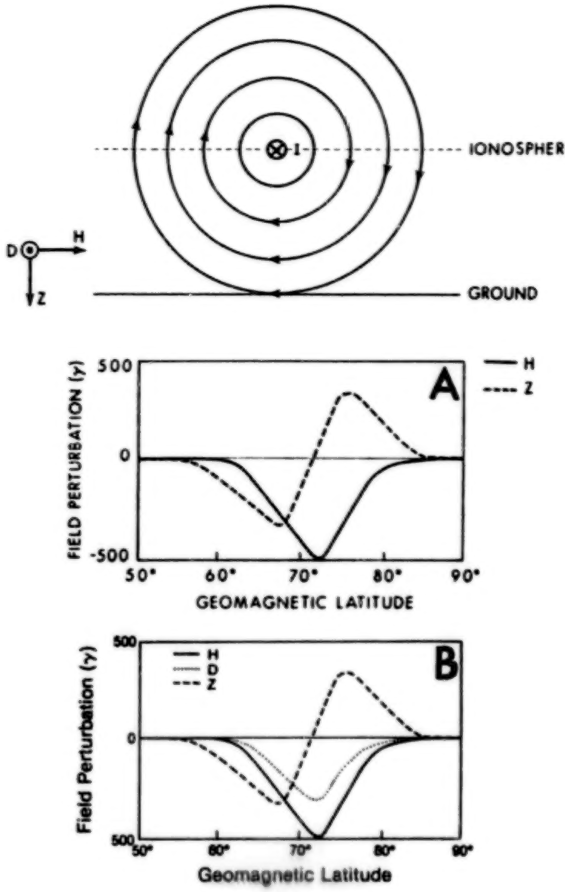


Figure 8. Panel A shows a cartoon of the latitude profile of magnetic disturbance due to an infinite westward ionospheric line current flowing normal to the line of recording magnetometers (Rostoker, 1972b). Panel B shows a similar profile where the current flows at an angle to the normal to the magnetometer line.

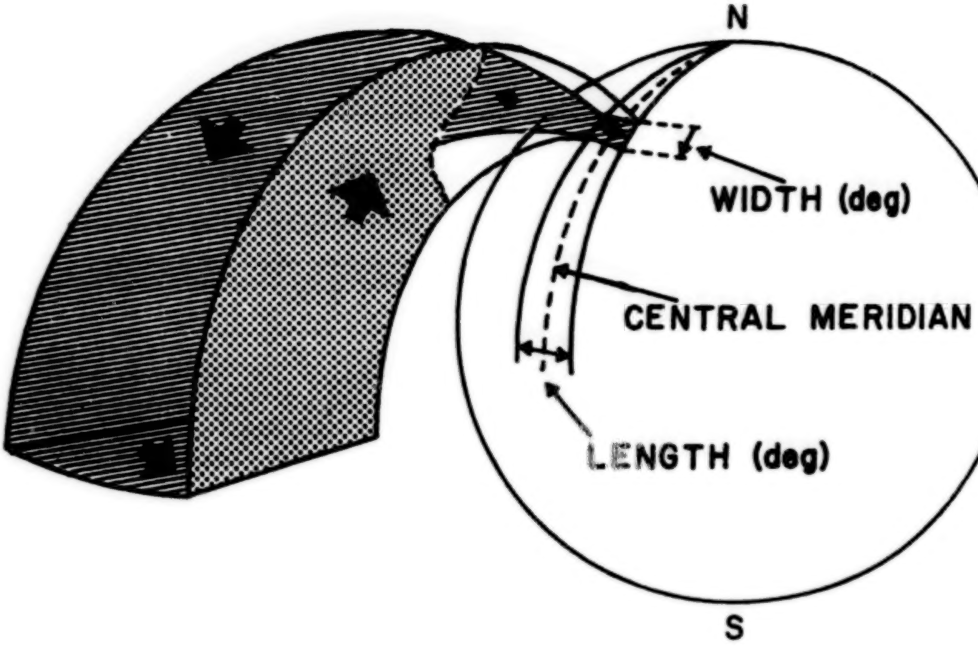


Figure 9. Three-dimensional current system in dipole field geometry representative of the expected current flow associated with substorm expansive phase activity (Kisabeth and Rostoker, 1977).

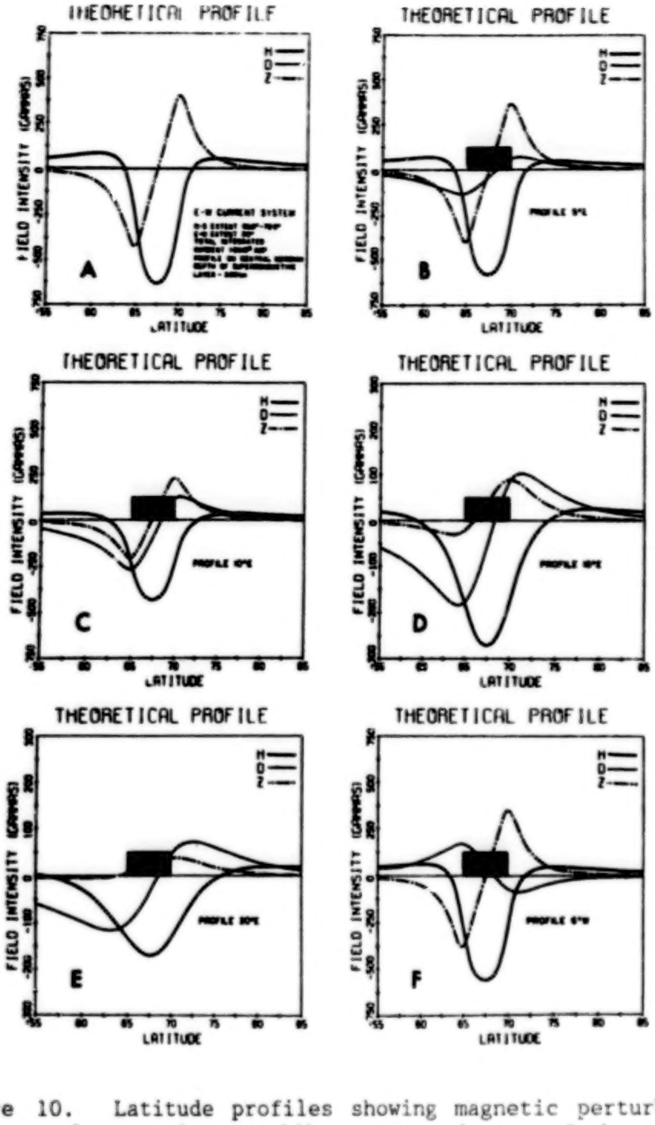


Figure 10. Latitude profiles showing magnetic perturbation patterns along various meridians east and west of the central meridian of the substorm westward electrojet (Kisabeth, 1972). These profiles include contributions due to induced currents in the Earth as described in the text.

Figure 11 shows the expected magnetic perturbation pattern along meridians in several representative magnetic local time sectors. These profiles were prepared by Hughes and Rostoker [1979] using meridian line data taken over several months in the northern hemisphere winter. Their preparation involved the identification of the center of the electrojet (viz. the latitude where ΔZ = 0) and performing the equivalent of a superposed epoch analyses using the center of the electrojet as the key location.

It is evident that the directly driven system involves a clear eastward electrojet across the dusk meridian and a westward electrojet across the dawn meridian as evidenced from the H-component and Z-component profiles (see Figure 8). The typical widths of those driven system electrojets is ~6° of latitude (viz. ~600-700 km) during average levels of activity. Near noon it is difficult to detect clear electrojet signatures while in the evening sector near midnight, one finds the signature of a westward electrojet current lying at the poleward edge of an eastward electrojet. (It is worth noting that the evening sector westward electrojet can experience large rapid transient changes in both width and magnitude. Then changes are associated with substorm expansive phase effects which will be discussed later in this paper.) While the H- and Z-components of the magnetic perturbation patterns shown in Figure 11 give vital information about overhead electrojet currents, the D-component gives much information about field-aligned currents penetrating the auroral oval and any azimuthal asymmetries in ionospheric current flow. In particular, the level shift in the D-component going from north to south in the noon sector indicates net downward field-aligned current flow across the local time regions, while the level shift in the D-component across the late evening sector indicates net upward field-aligned current. Current densities of these field-aligned currents have been established empirically by Hughes and Rostoker [1979] and can be evaluated for individual events through assumption of a reasonable forward model for the global current system involved in magnetosphere-ionosphere coupling (cf. Rostoker and Mareschal, 1981).

Finally, it should be noted that the net field-aligned current sheets described above create a distinctive magnetic signature over the polar cap region. In the early days of magnetospheric research, this magnetic signature was attributed to overhead electric current flow in the polar cap ionosphere. Ultimately, Heppner et al. [1971] were able to demonstrate that the polar cap magnetic signature was primarily due to the field-aligned current sheets which reflect the net upward and downward current flow across the evening and noon sectors respectively. This is an excellent example of a situation where equivalent current flow (across the polar cap) does not reflect the actual overhead flow of real electric current.

MAGNETIC SIGNATURES OF SUBSTORM EXPANSIVE PHASE ACTIVITY

The base substorm expansive phase current system is often called the "substorm current wedge" and is viewed as an ionospheric westward electrojet across the midnight region linked to the magnetosphere by current into the ionosphere at its eastern edge and out of the ionosphere at its western edge (cf. Figure 12). The magnetic signatures of this current loop as seen in the three components of the perturbation magnetic field are shown in magnetogram format in Figure 13 for various locations of observations on the nightside of the earth. In the context of this cartoon, the evolution of the current wedge involves spreading of the ionospheric electrojet both eastward and westward, although the westward spreading is viewed as dominant. Thus, the central meridian (bisecting the ionospheric electrojet) is found to move westward as

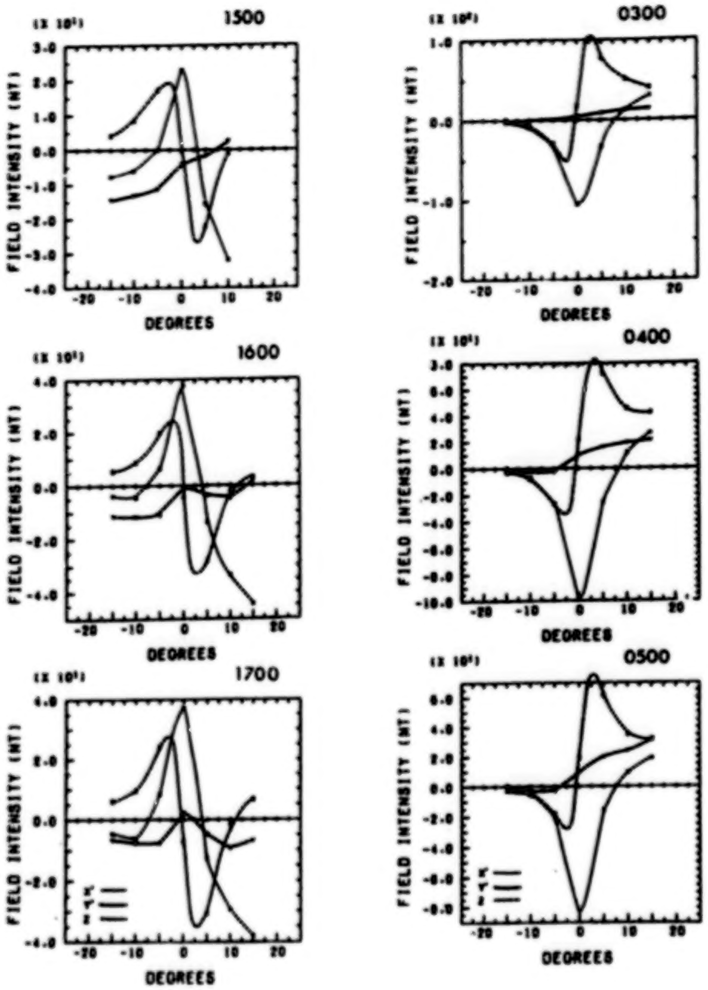


Figure 11. Averaged profiles in various magnetic local time sectors computed from three months of hourly averaged values from a meridian line of stations keyed to the center of the electrojet crossing the meridian. The panels on the left hand side of the figure are representative of afternoon sector profiles (MLT given at the top of each panel) while those on the right side of the figure are for the morning hours (Hughes and Rostoker, 1979).

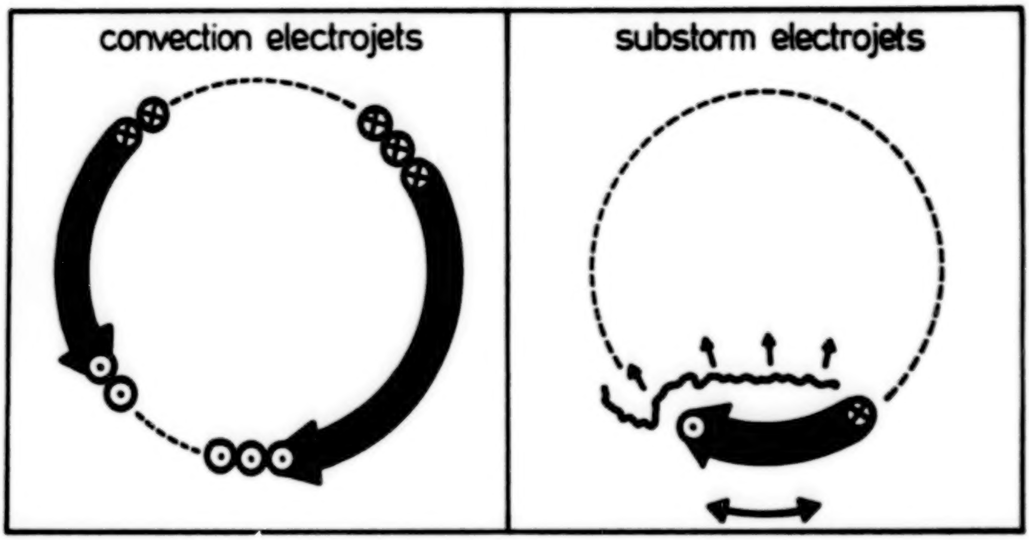


Figure 12. Cartoon representation of the substorm current wedge placed in the context of the directly driven auroral electrojets (Baumjohann, 1983).

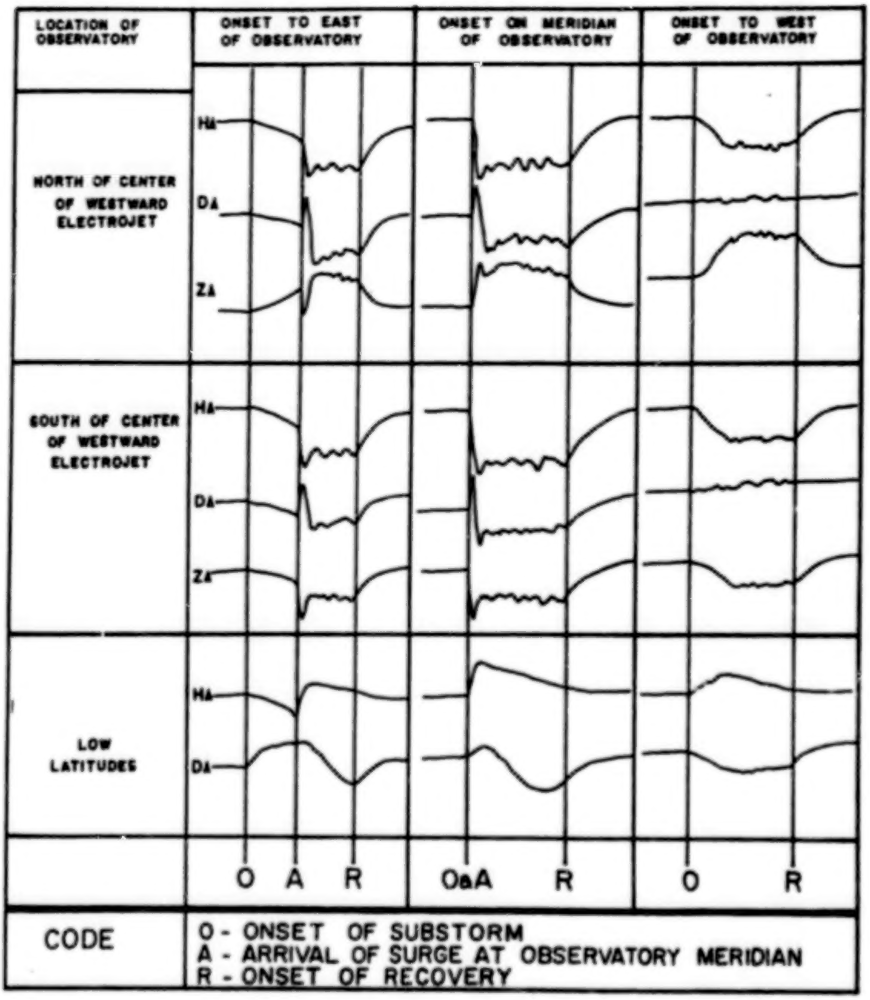

LOCATION OF OBSERVATORY
ONSET TO EAST OF OBSERVATORY
ONSET ON MERIDIAN OF OBSERVATORY
ONSET TO WEST OF OBSERVATORY
NORTH OF CENTER OF WESTWARD ELECTROJET
SOUTH OF CENTER OF WESTWARD ELECTROJET
LOW LATITUDES
HΔ
DΔ
ZΔ
O
A
R
O&A
CODE
O - ONSET OF SUBSTORM
A - ARRIVAL OF SURGE AT OBSERVATORY MERIDIAN
R - ONSET OF RECOVERY

Figure 13. Magnetic signatures of the development of the substorm current wedge at various locations on all sides of the current configuration. The placement of the auroral oval observatory closest to the local time of expansive phase onset is such that the initial surge formation is slightly to the east of the observation point (Rostoker et al., 1980).

the substorm expansive phase evolves (Rostoker et al., 1979) leading to transition D-component bays at subauroral latitudes (Rostoker, 1966) where the D-component deflects first positively and then negatively at an appropriately placed station sometime during the lifetime of the expansive phase.

While the picture of a monolithic wedge suffices for the interpretation of subauroral magnetograms, it is quite inadequate in permitting interpretation of ground magnetic data in the substorm-disturbed region of the auroral oval. This is because the overall substorm expansive phase current system evolves through the appearance of many small scale filamentary current systems whose combined magnetic perturbations yield the impression of a large scale current wedge when detected some distance from the disturbed region. Figure 14 shows, schematically, a typical distribution of these localized current loops after several minutes of expansive phase evolution. Each current element forms, normally, poleward of the previous one and with its western edge further to the west of the previous one. The western edge of each current filament often features a distorted auroral form called a surge. Figure 15 shows the latitude profiles during the evolution of a large substorm expansive phase. These profiles clearly show the poleward expansion of the substorm disturbed region and the multiple peaks in the H-component are clear evidence of the multi-filament character of the evolving substorm westward electrojet.

Fig. 16 shows a magnetogram which features a single isolated substorm filament. This magnetogram yields some information on the lifetime of an individual filament (i.e. several minutes) and the magnitude of its associated magnetic perturbation (~150-200 nT in the H-component directly under the filament). The physics of the substorm expansive phase comes from the study of the development of the individual filamentary current systems rather than the large scale wedge structure (which reflects the overall magnetic perturbation pattern due to the several contributing, filamentary systems). The magnetic signature of an individual filament greatly resembles that of the larger scale current wedge, however for a single filament the magnetic perturbation is barely detectable further than a few degrees from the ionospheric current channel. (Subauroral stations detect substorm expansive phases only when several filaments co-exist, so that their magnetic effects are additive.) Close to a filament the overall magnetic perturbation is dominated by ionospheric current flow, viz. a latitudinally and longitudinally localized westward electrojet producing a negative H-component bay. Near the western edge of the filament, a circulating Hall current around the localized region of upward current flow produces the characteristic signature of a positive D-component perturbation at the western edge of the associated auroral surge and a negative D-component to the east of the center of the upward field-aligned current (Tighe and Rostoker, 1981). Positive D-component bays can, however, result from either growth of the upward field-aligned current (detectable up to hundreds of km from the region of ionospheric outflow) or from the equatorward ionospheric current flow at the western edge of the surge (normally dominating a region of a scale size of ~200 km or less). An absence of significant Z-component fluctuations generally indicates that a positive D-component bay is attributable to the distant effect of field-aligned currents rather than nearby equatorward ionospheric current flow.

Finally, it is useful to comment on the impulsive magnetic pulsations known as Pi 2's. Following the early study of Saito [1961], Pi 2's (or pt's as they were known at that time) were considered to accompany geomagnetic bays observed at subauroral latitudes on a one-to-one basis. Subsequently, Rostoker [1968] demonstrated that a bay-type disturbance might feature several Pi 2

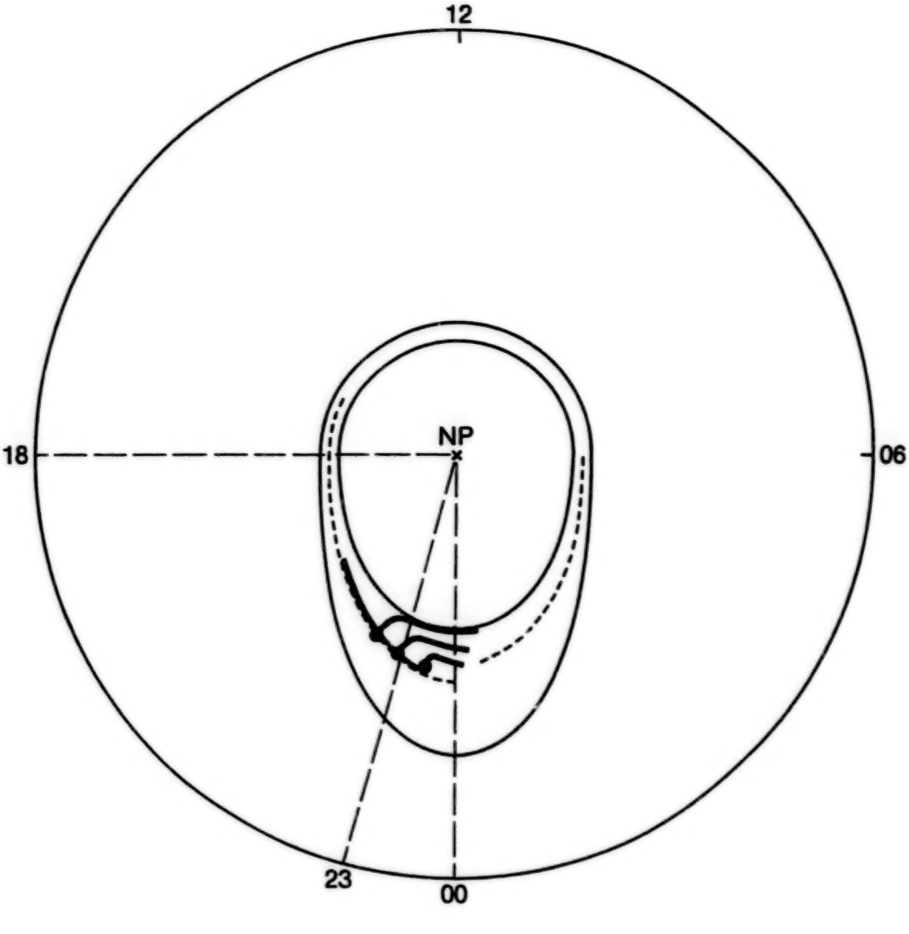


Figure 14. Cartoon showing the development of the filamentary current elements, the ensemble of which produce the magnetic signature of the substorm current wedge. Appropriately placed observatories yield magnetic signatures which clearly show the wedge to be composed of these small scale structures which tend to appear near midnight at expansive phase onset and are seen progressively westward and poleward as the expansive phase develops. The western edge of each structure is the site of upward field-aligned current while downward field-aligned current is distributed diffusely along the track of the westward electrojet element to the east.

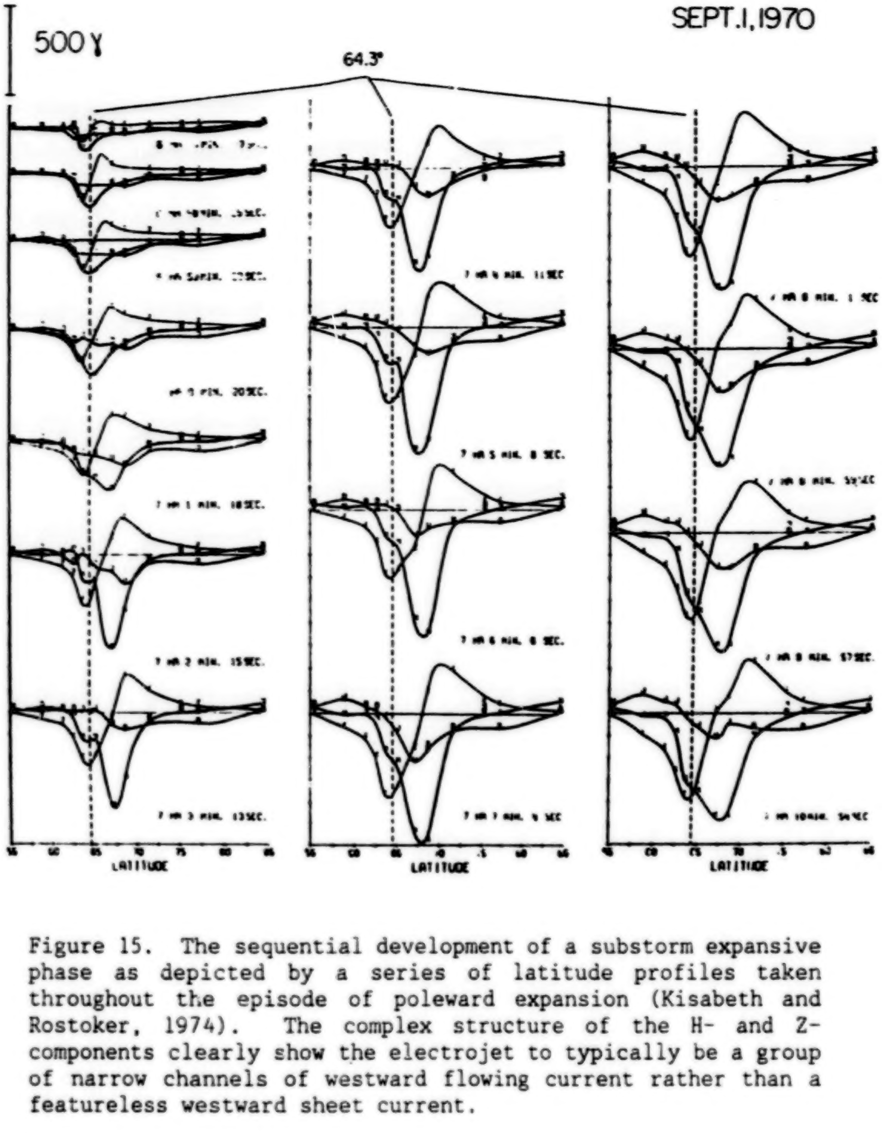


Figure 15. The sequential development of a substorm expansive phase as depicted by a series of latitude profiles taken throughout the episode of poleward expansion (Kisabeth and Rostoker, 1974). The complex structure of the H- and Z-components clearly show the electrojet to typically be a group of narrow channels of westward flowing current rather than a featureless westward sheet current.

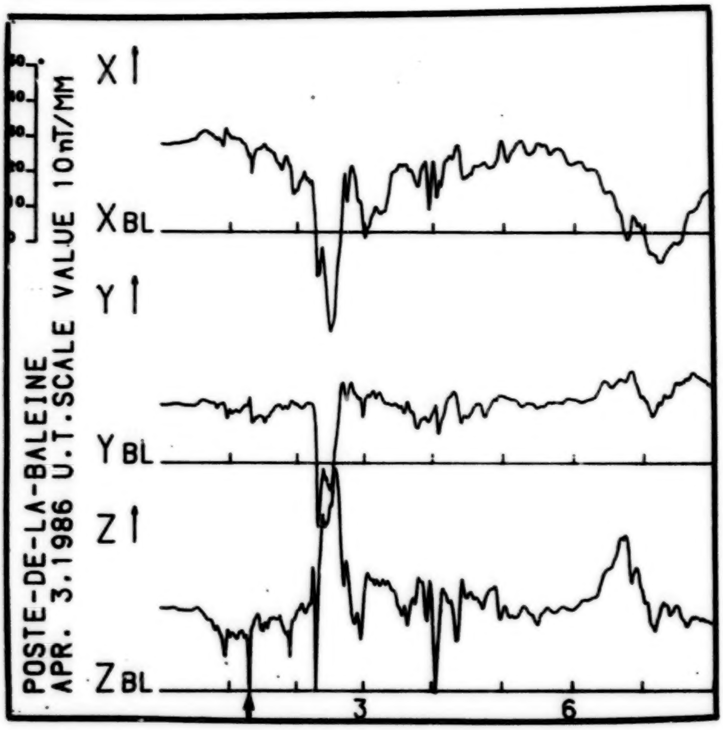


Figure 16. Magnetogram from an auroral oval observatory featuring the onset and evolution of a single filamentary structure at approximately 0117 UT on April 3, 1986. Many of these filaments occurring in close succession produce the overall signature of the substorm current wedge.

pulsation bursts, leading eventually to the concept of multiple onset substorms (cf. Wiens and Rostoker, 1975; Pytte et al., 1976). In recent times, it has become apparent that each expansive phase onset or intensification (featuring the development of a localized current filament as discussed above) involves a characteristic Pi 2 wavelet as shown in Figure 17a, with a multiple onset substorm involving many such wavelets which, when strung together with appropriate phasing, provide the appearance of a pulsation wave train (Figure 17b). The amplitudes of Pi 2 wavelets can reach a few tens of nT in the near vicinity of filament formation, however amplitudes a fraction of an nT up to a few nT are more typical for stations more than ~ 100 km away from the source region. Pi 2 pulsations are often used to time the onset of a substorm expansive phase or its subsequent intensifications. In this respect, it is important that researchers do not set some arbitrary magnitude threshold at single stations on which they base the conclusion that an onset/intensification did or did not occur. The onset of a clear Pi 2 wavelet clearly defines the time of the explosive growth of a substorm expansive phase current filament. However, the absence of a clearly defined Pi2 wavelet may merely indicate that the source was well removed from the available observation point(s).

Finally, it is interesting to note that there may even be structure in the filamentary currents whose appearance marks expansive phase onsets or intensifications. However, the scale size of such structures is actually <100 km and, since ground magnetometers are located at distances in excess of 100 km from the source current systems, it is impossible for ground based magnetometer arrays to study the evolution of current structure within the expansive phase filaments.

INTERPRETATION OF LATITUDE AND LONGITUDE PROFILES OF GROUND BASED MAGNETOMETER DATA

Earlier in this paper, we have shown in Figure 8 the expected magnetic perturbation pattern for an east-west ionospheric electrojet. Under the assumption that the electrojet is of infinite length, the perturbations only appear in the H- and Z- components. The eastward and westward electrojets due to directly driven activity are both very long and can be treated as infinite sheet currents of limited latitudinal width for purposes of identifying electrojet width and strength in the local time sector of the observations. Figure 18 shows a latitude profile taken in the afternoon sector featuring a clear eastward electrojet. Using the quantitative modeling techniques developed by Kisabeth [1972], it can be shown that the electrojet lies between the positive and negative extrema in the Z-component profile. Those same modeling techniques can be used to evaluate the strength of the current flowing in the electrojet. What is noticeable in Figure 18 is the level shift in the D-component across the latitude span of the electrojet. Based on the results of Hughes and Rostoker [1977, 1979], this level shift is interpretable in terms of a sheet of upward field-aligned current flow spanning the poleward portion of the electrojet. While the D-component carries information about the net upward or downward field-aligned current integrated along the meridian of the observations, that same component can also give information about azimuthal asymmetries in the current flow along the auroral oval. This is particularly valuable for the study of auroral surges and omega bands which accompany magnetospheric substorm activity. For studies of localized current structures within the oval, it is useful to have profiles of the magnetic perturbations along lines of both constant magnetic latitude and longitude. To give some indication of how one goes about interpreting these longitude and latitude

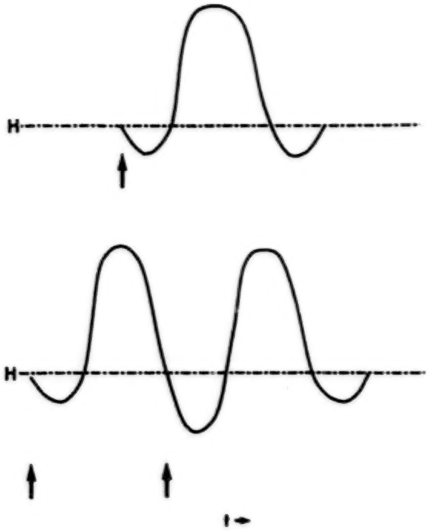


Figure 17. Panel A shows the magnetic signature of a Pi 2 pulsation which accompanies the onset of a single filamentary current system such as that shown in Figure 16. Panel B shows that the overall Pi 2 activity accompanying a multiple onset substorm consists of a number of these wavelets strung together. If the phasing of the wavelets is appropriate, the effect is that of a lengthy Pi 2 wavetrain as in the case of the two wavelets whose onsets (vertical arrows) are shown in Panel B.

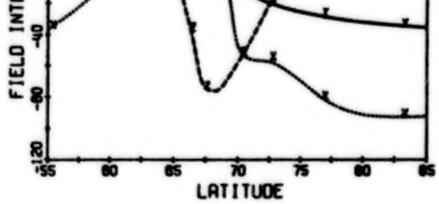


Figure 18. Latitude profile taken by the University of Alberta meridian line of magnetometers across the eastward electrojet in the early evening sector (~2000 MLT) on January 22, 1972. The electrojet current is distributed between the extrema in the Z-component profile while the negative level shift in the D-component from 67-70° is indicative of the presence of net upward field-aligned current flow in an east-west aligned current sheet crossing the meridian line of magnetometers.

profiles, we show in Figures 19a and 19b profiles taken during the development of an auroral surge associated with a substorm expansive phase.

Figure 19a shows, in Panel A the latitude profile across the auroral oval just prior to the development of a surge in the vicinity of the magnetometer array. The signature is that of a weak westward electrojet crossing the meridian line between 65° and 68° N. Panels B1, B2 and C in the same figure show the distortion of the magnetic perturbation pattern as the surge develops. The latitudinally localized positive D-component regime in Panel B1 is the signature of the localized region of equatorward current flow along the western edge of the surge form. The marked distortions in the H- and Z-components along the meridian line make their interpretation difficult based on Figure 19a alone. Figure 19b shows the corresponding data for the east-west line of stations (the north-south and east-west lines together forming a cross). Just prior to the appearance of the surge, Panel A shows H- and D-component signatures of a westward electrojet that changes its strength but little over the approximately one time zone of coverage. The increasing positive Z from west to east indicates that the electrojet is aligned from south-east to north-west, with the station to the north west being relatively close to the center of the electrojet. Panels B1, B2 and C show the longitudinal localization of the positive D-component regime associated with the surge. The movement of the positive peak in ΔD along with its zero crossover point and the negative peak in ΔZ are the signatures of the western expansion of the western edge of the surge as it evolves. The reader is referred to Tighe and Rostoker [1981] and Kawasaki and Rostoker [1979] for more detailed descriptions of the use of latitude and longitude profiles to diagnose the development of structures of localized east-west extent in the auroral oval.

EFFECTS OF INDUCED CURRENTS IN THE EARTH

No commentary on the interpretation of ground based data would be complete without a warning to the reader that any changes in the source current systems above the earth will induce electric currents in the conducting earth below the observation point (on or above the earth's surface). Thus the magnetic field perturbations recorded by a ground based magnetometer reflect the sum of the contributions from the source currents and the induced earth currents. In principle, it is possible to construct integral expressions for the separation of the internal and external contributions to the total perturbation magnetic field (Porath et al., 1970). However, in reality there is insufficient data available to permit such a separation of fields so that treatment of a time dependent problem involving finite subsurface conductivities is not viable. However, there have been several solutions proposed to the time-independent problem where the Earth is considered to be an insulator from the surface to some specified depth and a perfect conductor below that depth (cf. Bonnevier et al., 1970; Ashour, 1971; Kisabeth and Rostoker, 1977). For the purposes of studying three dimensional current systems of the type flowing during substorm activity, the approaches of Bonnevier et al. and Kisabeth and Rostoker are the most effective. In the former case, the authors considered each three dimensional current loop to be the sum of infinitesimal current loops around each surface element such that, if they are added together, the current cancels everywhere except at the border of the surface (see Figure 20). In this fashion, the magnetic perturbation from the three dimensional current loop can be considered as the sum of the contributions from a distribution of magnetic dipoles whose moments

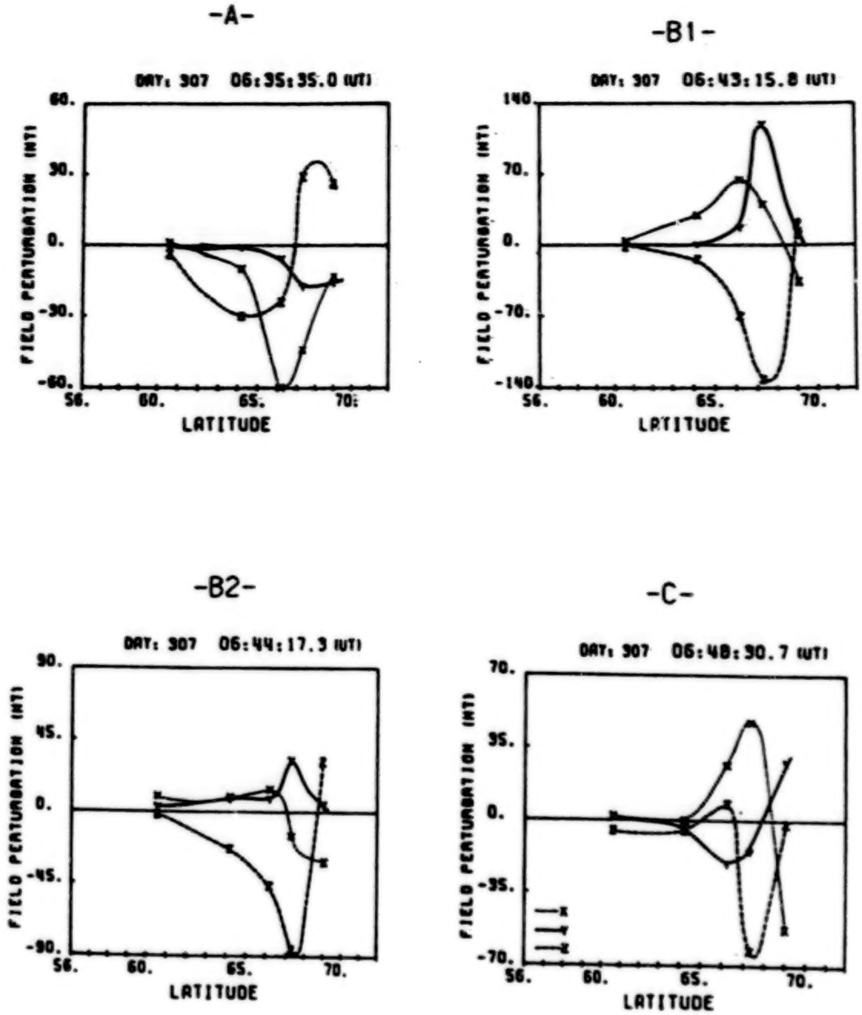


Figure 19a. Latitude profiles from a cross array showing the electrojet configuration just prior to the development of a surge near the meridian line (Panel A) and at three instants after the surge begins to develop (Panels B1, B2 and C) (Tighe and Rostoker, 1981).

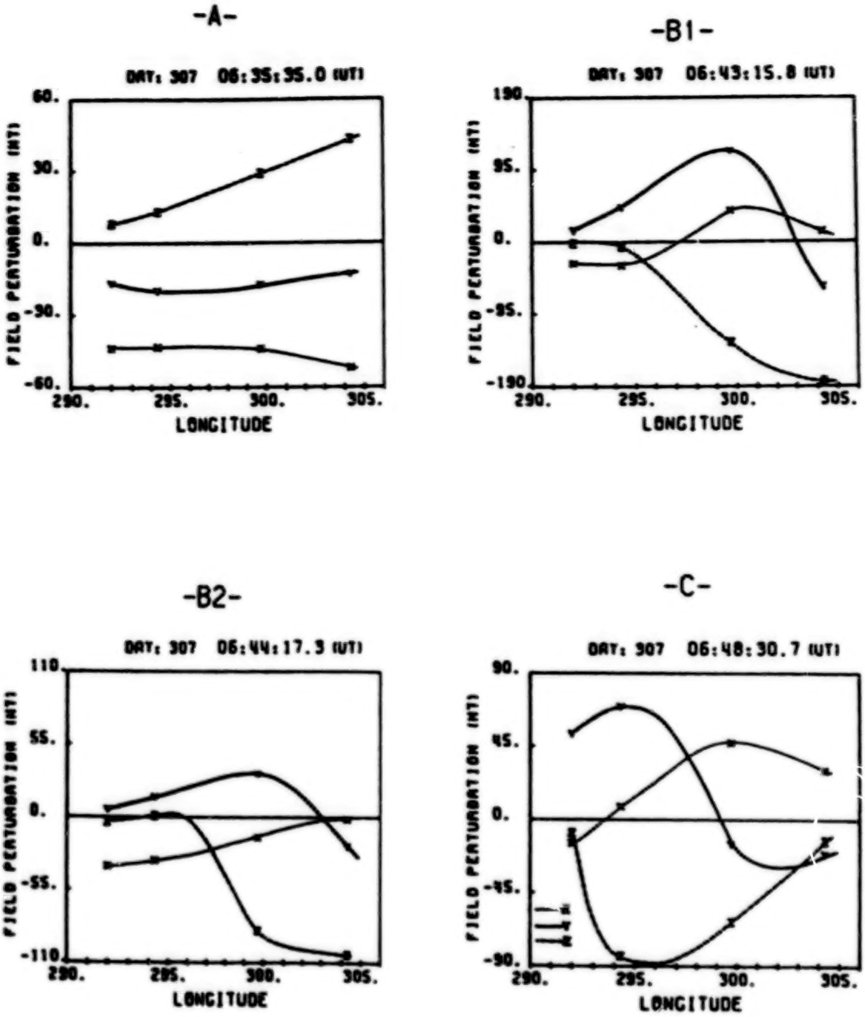


Figure 19b. Longitude profiles from the cross array alluded to in Figure 19a for equivalent moments during the substorm evolution described by the latitude profiles shown in that Figure (Tighe and Rostoker, 1981).

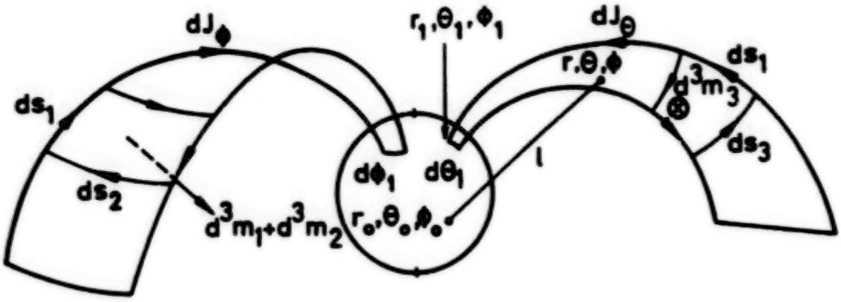


Figure 20. Geometry of the dipole current elements used to compute the magnetic effects of the three dimensional current loops descriptive of the substorm current wedge (Bonnevier et al., 1970).

$$\underline{d^3m} = d^3m_1\,\hat{\underline{r}} + d^3m_2\hat{\underline{\Theta}} + d^3m_3\hat{\underline{\phi}} \qquad [4]$$

are given by the size and orientation of the surface elements and the current in the loops. If one then considers the Earth to be represented by a perfect conductor from the center to radius r_2 and an insulator from r_2 to the surface, then the magnetic effect of the induced currents can be represented by image dipoles located within the conductor. For the radial component of the external dipole contribution (cf. Eq. 4), one obtains an image dipole with a moment $(-\ r_2^3/r^3)\ \underline{d^3m}_1$ located at the inverse point of the external dipole (i.e. at a distance $r^1 = r_2^2/r$ from the center). For $\underline{d^3m}_2$ and $\underline{d^3m}_3$, one gets similar image dipoles parallel to the external ones at the inverse point plus a dipole distribution along a straight line between the center $x = 0$ and the inverse point $x = r^1$ with the moments $-\ (x\ dx/r\ r_2)\ [\underline{d^3m}_2 + \underline{d^3m}_3]$. The total disturbance field is then obtained by adding the contributions of the external and image dipoles at each field point where the perturbation field is desired. As an example of the influence of induced currents in the Earth, we show in Table 1 the ratio of internal to external field contributions at various latitudes for a three dimensional current loop involving a westward electrojet of length 20° located with its center at 67.5°N and having a width of 5° of latitude. The ratios shown in Table 1 are calculated for depths of the perfect conductor ranging from 100 km to 300 km. This is because, in line with the expectation for skin depths of the real Earth, we should expect the perfect conductor to be close to the surface for high frequency disturbances and deeper for low frequency disturbances. For disturbances having the time scale of substorms, Kisabeth [1972] has found that 250 km is the most appropriate depth for the surface of the perfect conductor. The effect of subsurface conductivity is to increase the magnitude of the horizontal component of the magnetic perturbation while decreasing the vertical component. For a flat earth approximation and an infinite line current source, the correction factors are 2/3 and 2 for the horizontal and vertical components respectively. The ratios shown in Table 1 clearly show that such a uniform treatment for all field points is not justified, and it is necessary to know the distance of the observation point from the source current distribution as well as the geometry of the source current distribution if one is to effectively correct for Earth induction effects.

Finally, it is important to note that the treatment of the induction problem shown above is only appropriate if there are no regions of anomalous subsurface conductivity in the vicinity of the observation points. Unfortunately, the Earth is far from having a uniform conductivity and sharp discontinuities are often found near tectonically active regions in addition to regions that have been tectonically active in the distant past. As well, the effects of seawater together with the upwelling of the Earth's mantle at continental margins make rather suspect the interpretation of magnetic perturbations from substorms detected by stations near coastlines. The reader is referred to Rikitake [1966] for a more detailed description of the effects of subsurface conductivity anomalies on the perturbation magnetic field detected at the Earth's surface together with the analytical techniques one can use to establish the presence of such anomalies.

INFERENCE OF MAGNETOSPHERIC AND IONOSPHERIC ELECTROMAGNETIC PROPERTIES FROM QUANTITATIVE MODELING USING GROUND BASED MAGNETOMETER ARRAY DATA

Up till now, this paper has concentrated on providing a qualitative description of how one goes about interpreting the signatures found in ground magnetometer data. However, one should hope that such information can provide a more quantitative assessment of the spatial extent of the disturbed regions

Table 1: Ratio of the total perturbation magnetic field due to internal induced currents to the total perturbation magnetic field due to external inducing currents from a three dimensional current system (Figure 9) for different observation points on the surface of the perfect conductor below the Earth's surface.

Latitude	100 km	200 km	300 km
55°	0.836	0.691	0.560
60°	0.736	0.526	0.375
65°	0.416	0.221	0.133
67.5°	0.452	0.238	0.142
70°	0.400	0.210	0.127
75°	0.751	0.532	0.377
80°	0.874	0.725	0.590
84.5°	0.924	0.816	0.709

of the ionosphere and the possibility of estimating some of the more crucial parameters through which one might hope to quantify the level of the solar terrestrial interaction. During the early part of the 1970's, analytical routines were developed which permitted the magnetic perturbation from three dimensional current systems to be evaluated at any point on the Earth's surface (cf. Bonnevier et al., 1970; Kisabeth and Rostoker, 1977; Kisabeth, 1979). While these calculations were relevant to current flow in a dipolar magnetic field geometry, the fact that most of the perturbation detected on the surface of the Earth comes from current flow relatively close to the Earth made such model calculations rather useful. However, in order to achieve further progress, it is necessary to have the current flow set in a more realistic geometry involving high latitude currents flowing out into the deep magnetotail instead of being confined to dipole geometry. For this reason, we will not develop the codes used to compute perturbations from three dimensional current systems in a dipole magnetic field in this paper. The reader can find the details in the papers quoted above along with full comparisons of the various approaches to the problem in Kisabeth [1972].

By the beginning of the 1980's, a new approach to modeling was coming to the fore. This method emphasized the physics of ionospheric current flow and the reasons for the divergence of this flow along magnetic field lines so as to produce the three dimensional current loops which couple the ionosphere and magnetosphere. Development of the codes to implement this modeling scheme proceeded independently in the U.S.S.R. (cf. Bazarzhapov et al., 1979; Mishin et al., 1980) and in the U.S.A. (cf. Kamide and Matsushita, 1979a, 1979b; Kamide et al., 1981) leading to the ability to infer information about the distribution of ionospheric currents, field-aligned currents and electric fields in the high latitude regions of the Earth. In the light of the importance of the development of such algorithms for the effective utilization of ground based magnetometer observations, we shall briefly outline the basis of this modeling approach as developed by Kamide et al. [1981] (known as the KRM technique).

One begins by utilizing the world wide network of magnetometers to establish the equivalent current system at some point in time. Here the horizontal magnetic perturbation vectors are computed (subtracting the quiet time values from the measured values and making some effort to correct for Earth induction effects) and the equivalent current system constructed by drawing smooth curves of equivalent current flow. If one describes the equivalent current flow lines by the scalar potential function Ψ, the equivalent current density (in A/m) can be expressed in the form

$$\underline{j}_T = \underline{n}_r \times \text{grad}\ \Psi \qquad [5]$$

where $\underline{n}_r$ is a unit radial vector. Outside of the region of real current flow, the measured magnetic perturbation can be expressed as the gradient of a magnetic scalar potential, viz.

$$\underline{b} = -\ \text{grad}\ V \qquad [6]$$

and Chapman and Bartels [1940] (pg. 631) have shown that the external portion of V is related to Ψ by straightforward mathematical expressions.

The KRM technique proceeds on the basis that, once one knows the equivalent current function Ψ, one can assume a reasonable conductivity distribution (for which one must have an analytical expression) and derive the ionospheric electric field, the real horizontal current and the field-aligned current under the following set of assumptions:

(1) The electric field is purely electrostatic.

(2) Geomagnetic field lines are electric equipotentials.

(3) The dynamo effect of neutral winds can be ignored.

(4) Magnetic contributions due to magnetospheric ring currents, tail currents and magnetopause currents can be ignored.

(5) Geomagnetic field lines are effectively radial, implying that errors due to the fact that field-aligned currents actually follow the quasi-dipolar magnetic field lines can be ignored.

The next step involves the expression of the total height-integrated horizontal ionospheric current as the sum of the equivalent current and a "potential" current $\underline{j}_P$, viz.

$$\underline{j} = \underline{j}_T + \underline{j}_P \qquad [7]$$

where $\underline{j}_P$ can be expressed in terms of a current potential τ, viz.

$$\underline{j}_P = - \text{grad } \tau \qquad [8]$$

such that div $\underline{j}_T = 0$ and curl $\underline{j}_P = 0$ and the field aligned-current density is given by

$$j_{||} = \text{div } \underline{j} = \text{div } \underline{j}_P \qquad [9]$$

where $j_{||}$ is positive downwards.

The real horizontal ionospheric current is related to the electric field $\underline{E}$ (in a frame of reference corotating with the Earth) through the expression

$$\underline{j} = \Sigma_P \underline{E} + \Sigma_H \, \underline{E} \times \underline{n}_r \qquad [10]$$

where

$$\Sigma_P = \int_o^h \sigma_P \, dr$$

and

$$\Sigma_H = \int_o^h \sigma_H \, dr$$

are the height-integrated Pedersen and Hall conductivities respectively, with σ_P and σ_H being the respective specific conductivities. Finally, one expresses the electric field in terms of scalar potential

$$\underline{E} = - \text{grad } \Phi \qquad [11]$$

and obtains a partial differential equation for Φ in terms of Ψ by equating the right hand sides of Eq (7) and Eq (10) and taking the curl of the resulting expression. In spherical coordinates, where θ is the colatitude and λ is the east longitude, one obtains

$$A \frac{\partial^2 \Phi}{\partial \theta^2} + B \frac{\partial \Phi}{\partial \theta} + C \frac{\partial^2 \Phi}{\partial \lambda^2} + D \frac{\partial \Phi}{\partial \lambda} = F \qquad [12]$$

where the coefficients in the above equation are given by

$$A = (\sin\theta)\Sigma_H$$

$$B = \frac{\partial}{\partial \theta}((\sin\theta)\Sigma_H) + \frac{\partial}{\partial \lambda}\Sigma_P$$

$$C = \Sigma_H/\sin\theta$$

$$D = \frac{\partial}{\partial \theta}\Sigma_P - \frac{\partial}{\partial \lambda}(\Sigma_H/\sin\theta)$$

$$F = \frac{\partial}{\partial \theta}(\sin\theta \frac{\partial \Psi}{\partial \theta}) + (1/\sin\theta)\frac{\partial^2 \Psi}{\partial \lambda^2}$$

Eq (12) is solved using the boundary conditions

$$\Phi\,(0,\lambda) = 0 \qquad (13)$$

$$\frac{\partial \Phi}{\partial \theta}(\pi/2,\lambda) = 0 \qquad (14)$$

where the equatorial boundary condition expressed in Eq (14) is in place only for numerical convenience and has no physical basis in fact. A finite difference scheme is then used to solve the partial differential Eq (12) yielding the form of the electrostatic potential Φ. The ionospheric electric field can then be obtained from

$$E_\theta = -\frac{\partial \Phi}{a\,\partial \theta} \qquad E_\lambda = -\frac{\partial \Phi}{a \sin\theta\,\partial \lambda} \qquad (15)$$

where a is the distance from the center of the earth to the (thin) ionosphere. Using the values of E_θ and E_λ and reformulating the height-integrated conductivities in the (λ,θ) coordinate system, viz.

$$\underline{\underline{\Sigma}} = \begin{bmatrix} \Sigma_P & \Sigma_H \\ -\Sigma_H & \Sigma_P \end{bmatrix} = \begin{bmatrix} \Sigma_{\theta\theta} & \Sigma_{\theta\lambda} \\ -\Sigma_{\lambda\theta} & \Sigma_{\lambda\lambda} \end{bmatrix}$$

the ionospheric current can be expressed in the form

$$\begin{bmatrix} j_\theta \\ j_\lambda \end{bmatrix} = \begin{bmatrix} \Sigma_{\theta\theta} & \Sigma_{\theta\lambda} \\ -\Sigma_{\theta\lambda} & \Sigma_{\lambda\lambda} \end{bmatrix} \begin{bmatrix} E_\theta \\ E_\lambda \end{bmatrix} \qquad (16)$$

Finally, by inserting the information about the real ionospheric current from Eq (16) into Eq (9), one can obtain the field-aligned current. Kamide et al. used the accelerated Liebmann over-relaxation method (Gary, 1969) involving an iteration technique to minimize a convergence parameter in order to evaluate the current and electric field values over the grid of field points.

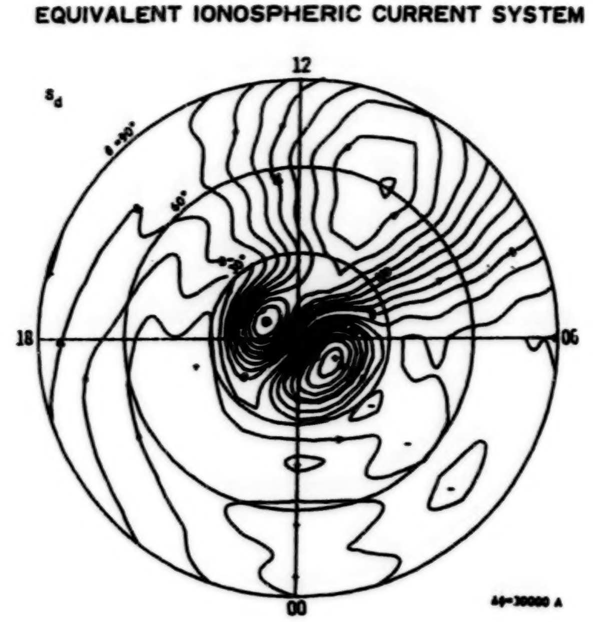
To demonstrate the capability of the KRM algorithm, we show in Figure 21 the output for a run performed for ground data obtained over the interval May-June 1965 during disturbed days. The height-integrated conductivities due to energetic particle precipitation were approximated by the expressions

$$\Sigma_P = \Sigma_m \exp\left[-\frac{(\theta-\theta_o)^2}{D_\theta{}^2} - \frac{(\lambda-\lambda_o)^2}{D_\lambda{}^2}\right]$$

$$\Sigma_H = k\Sigma_P \tag{17}$$

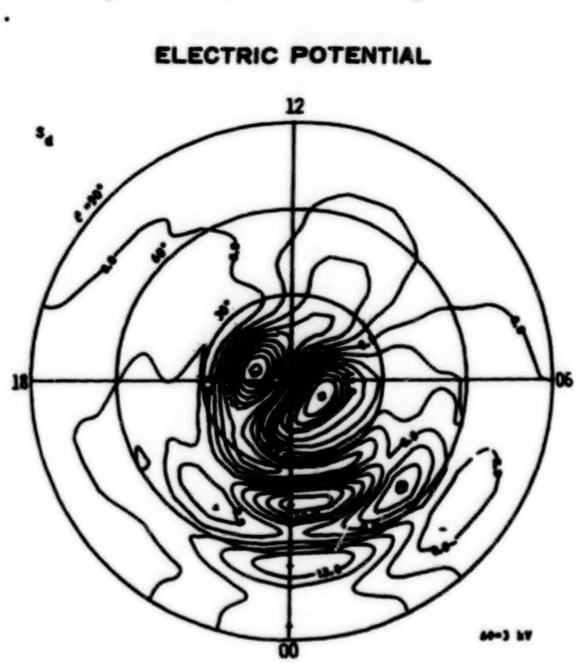
and these values were superposed on an appropriate conductivity distribution due to the incidence of solar radiation on the ionosphere. In Eq (17), (θ_o,λ_o) gives the position of the center of the region of enhanced conductivity while D_θ and D_λ specify the Gaussian distribution for the latitudinal and longitudinal directions respectively with Σ_m and k being constants. Figure 21a shows the equivalent current system for the interval in question as a polar projection in geomagnetic coordinates; this information constituted the input to the routine obtained using ground based magnetometer data. Figure 21b shows the equipotential distribution in the same polar plot format, this figure containing the calculated information for the ionospheric electric field. Figure 21c shows the real ionospheric current vectors for latitudes higher than 50° in which the eastward and westward directly driven electrojets are clearly evident. Finally, Figure 21d gives the field-aligned current distribution, in which the anti parallel pairs of current sheets discovered by Zmuda and Armstrong [1974] are apparent.

Ultimately, if one were to have an ideal two dimensional grid of magnetometer station covering the Earth's surface (grid spacing 1° x 1°) it would be possible to obtain a very accurate picture of the key electric field and current parameters, limited only by the inaccuracies of the conductivity distribution. However, in reality the available magnetometer stations are quite limited in number and irregularly spaced over the Earth's surface. Using only information about the horizontal components of the perturbation magnetic field, it is quite possible to be significantly in error in one's estimate of the current and field parameters. For example, if one were to have only one station in a given longitudinal sector and the station were to fortuitously lie at the interface between the eastward and westward electrojets in the evening sector, a zero horizontal perturbation magnetic field could be measured leading to the incorrect conclusion that no electrojet currents crossed the meridian of the station. Such a false conclusion would be used by the algorithm to produce entirely meaningless results regarding the currents and electric field anywhere in the vicinity of the unfortunately placed station. It is possible to partly alleviate this problem by using information from the Z-components of the magnetic perturbations at the many stations in the array. Combined with a reasonable forward model of the electrojet structure across the meridian of each observatory, it is possible to use information about the Z/H ratio to infer



EQUIVALENT IONOSPHERIC CURRENT SYSTEM

Figure 21a. Equivalent current system for disturbed days calculated by averaging data over the interval May - June 1965. This set of data obtained from ground magnetometer recordings constitutes the input for the KRM modeling routine (Kamide et al., 1981).



ELECTRIC POTENTIAL

Figure 21b. Equipotential distribution calculated using the KRM algorithm for the data portrayed in Figure 21a (Kamide et al., 1981).

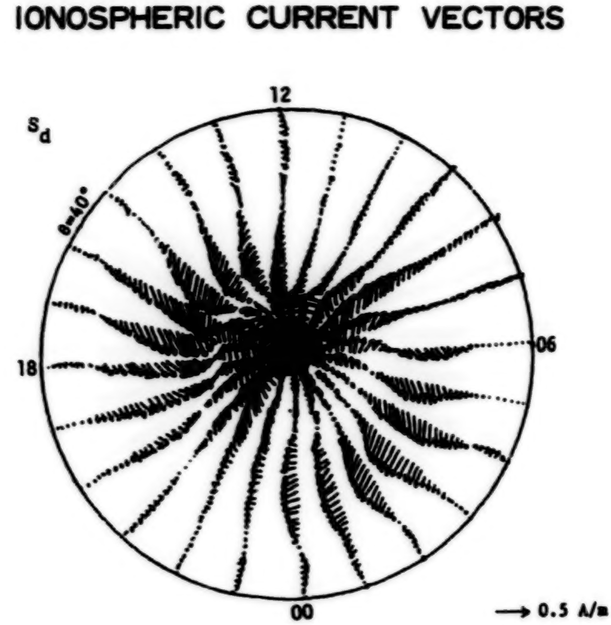


Figure 21c. Real ionospheric current vectors calculated using the KRM algorithm for the data portrayed in Figure 21a (Kamide et al., 1981).

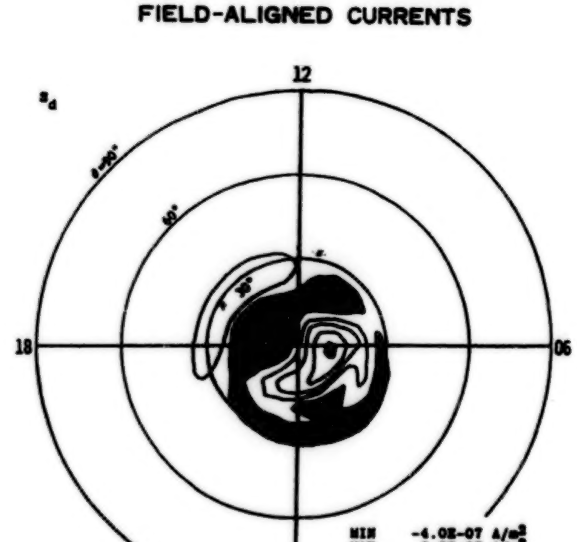


Figure 21d. Field-aligned current distribution calculated using the KRM algorithm for the data portrayed in Figure 21a (Kamide et al., 1981). The shaded region indicates upward current flow.

information about the maximum H-component perturbation along the meridian passing through the observatory. In some cases, it might be possible to use linear inversion techniques (Hughes et al., 1979) coupled with an appropriate forward model of the electrojet structure to accurately quantify the strength of the ionospheric current flow across a meridian along which observatories were arrayed with reasonable latitudinal separation.

Ultimately, it may be possible to use luminosity distributions obtained from imagers aboard high altitude satellites to infer useful information about the conductivity distribution in the high latitude ionosphere (cf. Marklund et al., 1987). Since such luminosity information might also yield the position of the poleward and equatorward borders of the auroral current systems, it may yet be possible to put additional constraints on the KRM technique in such a manner as to produce yet more reliable information on the instantaneous current and electric field configuration in the high latitude ionosphere.

CONCLUDING REMARKS

In this paper, I have tried to demonstrate how ground based magnetometer data can be used to infer information about the three-dimensional electric current systems which couple the magnetosphere to the ionosphere. It must be emphasized that magnetometer data from stations confined to the Earth's surface cannot, by themselves, provide a non-unique representation of the electric current systems external to the Earth. However, combined with information obtained from polar orbiting satellites carrying particle and field detectors, it has been possible to construct reasonable forward models for the current systems. The interpretation of ground magnetometer data presented in this paper has drawn on the wealth of correlations described in the literature between ground magnetometer data and polar orbiting satellite data to give a useful tool to the student of the solar-terrestrial interaction. While there is still considerable research to be done in identifying the signatures of relatively small scale current systems involved with auroral forms, the reader should feel comfortable in using the techniques described in this paper for the study of directly driven magnetospheric activity as well as many aspects of substorm expansive phase behaviour.

ACKNOWLEDGEMENTS

This research was supported by the National Sciences and Engineering Research Council of Canada.

REFERENCES

Ashour, A. A., The evaluation of the field of the currents induced in the Earth by an external field whose distribution is known numerically, Radio Sci., 6, 171, 1971.

Axford, W. I. and C. O Hines, A unifying theory of high-latitude geophysical phenomena and geomagnetic storms, Can. J. Phys., 39, 1443, 1961.

Bannister, J. R. and D. I. Gough, Development of a polar magnetic substorm: a two dimensional magnetometer array study, Geophys. J. R. Astr. Soc., 51, 75, 1977.

Bonnevier, B., R. Boström; and G. Rostoker, A three dimensional model current system for polar magnetic substorms, J. Geophys. Res., 75, 107, 1970.

Baumjohann, W., Ionospheric and field-aligned current systems in the auroral zone: a concise review, in Advances in Space Research., 2, 55, 1983.

Bazarzhapov, A. D., V. M. Mishin, D. S. Shirapov and G. B. Shpynev, Electric fields and currents in the quiet magnetosphere as determined from ground measurements, Issle. Geomag. Aeron. Fiz. Solntsa, 46, 13, 1979.

Chapman, S. and J. Bartels, Geomagnetism, Oxford Univ. Press (Clarendon), London, 1940.

Dungey, J. W., Interplanetary magnetic field and the auroral zones, Phys. Rev. Lett., 6, 47, 1961.

Eastman, T. E., L. A. Frank and C. Y. Huang, The boundary layers as the primary transport regions of the earth's magnetotail. J. Geophys. Res., 90, 9541, 1985.

Fukushima, N., Polar magnetic storms and geomagnetic bays, J. Fac. Sci. Tokyo Univ. 8, 293, 1953.

Gary, J., The numerical solution of partial differential equations, NCAR MS 69 - 54, Nat. Cent. for Atmosph. Res., Boulder, CO, 1969.

Goertz, C. K. and R. A. Smith, The thermal catastrophe model of substorms, J. Geophys. Res., 94, 6581, 1989.

Hakura, Y., Tables and maps of geomagnetic coordinates corrected by the higher order spherical harmonic terms, Rept. Ionos. Space Res., Japan 19, 121, 1965.

Harang, L., The mean field of disturbance of polar geomagnetic storms, Terrest. Mag. Atmos. Elec. 51, 353, 1946.

Heppner, J. P., J. D. Stolarik and E. M. Wescott, Electric field measurements and the identification of currents causing magnetic disturbances in the polar cap, J. Geophys. Res., 76, 6028, 1971.

Hines, C. O., The energization of plasma in the magnetosphere: hydromagnetic and particle-drift approaches, Planet. Space Sci., 10, 239, 1963.

Hughes, T. J., and G. Rostoker, Current flow in the magnetosphere and ionosphere during periods of moderate activity, J. Geophys. Res., 82, 2271, 1977.

Hughes, T. J. and G. Rostoker, A comprehensive model current system for high latitude magnetic activity - I. The steady state systems, Geophys. J. Roy. astron. Soc., 58, 525, 1979.

Hughes, T. J., D. W. Oldenburg and G. Rostoker, Interpretation of auroral oval equivalent current flow near dusk using inversion techniques, J. Geophys. Res., 84, 450, 1979.

Kamide, Y. and S. Matsushita, Simulation studies of ionospheric electric fields and currents in relation to field-aligned currents, 1. Quiet periods, J. Geophys. Res., 84, 4083, 1979a.

Kamide, Y. and S. Matsushita, Simulation studies of ionospheric electric fields and currents in relation to field-aligned currents, 2. Substorms, J. Geophys. Res., 84, 4099, 1979b.

Kamide, Y., A. D. Richmond and S. Matsushita, Estimation of ionospheric electric field, ionospheric currents, and field aligned currents from ground magnetic records, J. Geophys. Res., 86, 801, 1981.

Kamide, Y., B.-H. Ahn, S. -I Akasofu, W. Baumjohann, E. Friis-Christensen, H. W. Kroehl, H. Maurer, A. D. Richmond, G. Rostoker, R. W. Spiro, J. K. Walker and A. N. Zaitzev, Global distribution of ionospheric and field-aligned currents during substorms determined with magnetic data from six IMS meridan chains: initial results, J. Geophys. Res., 87, 8228, 1982.

Kawasaki, K. and G. Rostoker, Perturbations magnetic fields and current systems associated with eastward drifting auroral structures, J. Geophys. Res., 84, 1464, 1979.

Kisabeth, J. L., The dynamical development of the polar electrojets., Ph.D. Thesis, University of Alberta, Edmonton, Alberta, Canada 1972.

Kisabeth, J. L., On calculating magnetic and vector potential fields due to large-scale magnetospheric current systems and induced currents in an infinitely conducting earth, in: Quantitative Modeling of Magnetospheric Processes, ed. by W. P. Olson, American Geophysical Union, Washington, D. C., p. 473, 1979.

Kisabeth, J. L. and G. Rostoker, The expansive phase of magnetospheric substorms, 1. Development of the auroral electrojets and auroral arc configuration during a substorm, J. Geophys. Res., 79, 972, 1974.

Kisabeth, J. L. and G. Rostoker, Modelling of three-dimensional current systems associated with magnetospheric substorms, Geophys. J. Roy. Astron. Soc., 49, 655, 1977.

Lundin, R., On the magnetospheric boundary layer and solar wind energy transfer into the magnetosphere, Space. Sci. Rev., 48, 263, 1988.

Marklund, G. T., L. G. Blomberg, T. A. Potemra, J. S. Murphree, F. J. Rich and K. Stasiewicz, A new method to derive "instantaneous" high latitude potential distributions from satellite measurements including auroral imager data, Geophys. Res. Lett., 14, 439, 1987.

Mishin, V. M., A. D. Bazarzhapov and G. B. Shpynev, Electric fields and currents in the earth's magnetosphere, in Dynamics of the Magnetosphere, ed. by S. -I. Akasofu, p. 249, D. Reidel Publ. Co., Hingham, MA, 1980.

Perreault, P. and S. -I. Akasofu, A study of geomagnetic storms, Geophys. J. Roy. astron. Soc., 54, 547, 1978.

Porath, H., D. W. Oldenburg, and D. I. Gough, Separation of magnetic variation fields and conductive structures in the Western United States, Geophys. J. Roy. astr. Soc. 19, 237, 1970.

Pytte, T., R. L. McPherron and S. Kokubun, The ground signatures of the expansion phase during multiple onset substorms, Planet. Space Sci., 24, 1115, 1976.

Rikitake, T., Electromagnetism and the Earth's Interior, Elsevier Publ. Co., New York, 1966.

Rostoker, G., Midlatitude transition bays and their relation to the spatial movement of overhead current systems, J. Geophys. Res., 71, 79, 1966.

Rostoker, G., Macrostructure of geomagnetic bays, J. Geophys. Res., 73, 4217, 1968.

Rostoker, G., Geomagnetic indices, Rev. Geophys. Space Phys., 10, 935, 1972.

Rostoker, G., Interpretation of magnetic field variations during substorms, in Earth's Magnetospheric Processes, ed. by B. M. McCormac, p. 379, D. Reidel Publ. Co., 1972.

Rostoker, G., Overview of observations and models of auroral substorms, in Auroral Physics, ed. by C.-I. Meng, Cambridge Univ. Press, 1989, (in press).

Rostoker, G., C. W. Anderson III. D. W. Oldenburg, P. A. Camfield, D. I. Gough, and H. Porath, Development of a polar magnetic substorm current system, J. Geophys. Res., 75, 6318, 1970.

Rostoker, G. and R. Boström, A mechanism for driving the gross Birkeland current configuration in the auroral oval, J. Geophys. Res., 81, 235, 1976.
Rostoker, G. and M. Mareschal, Field-aligned current and the auroral electrojets in the post-noon quadrant, J. Geophys. Res., 87, 9071, 1982.
Rostoker, G., S. -I. Akasofu, J. Foster, R. A. Greenwald, Y. Kamide, K. Kawasaki, A.T.Y. Lui, R. L. McPherron and C. T. Russell, Magnetospheric substorms - definition and signatures, J. Geophys. Res., 85, 1663, 1980.
Saito, T., Oscillation of geomagnetic field with the progress of pt-type pulsation, Sci. Rept. Tohoku Univ., 5, Geophys. 13, 53, 1961.
Silsbee, H. C. and E. H. Vestine, Geomagnetic bays, their occurrence frequency and current systems, Terr. Magn. Atmos. Elec., 47, 195, 1942.
Tighe, W. G. and G. Rostoker, Characteristics of westward travelling surges during magnetospheric substorms, J. Geophys., 50, 51, 1981.
Wiens, R. G. and G. Rostoker, Characteristics of the development of the westward electrojet during the expansive phase of magnetospheric substorms, J. Geophys. Res., 80, 2109, 1975.
Zmuda, A. J. and J. C. Armstrong, The diurnal flow of field-aligned currents, J. Geophys. Res., 79, 4611, 1974.

Chapter 2

"IN SITU" MEASUREMENT TECHNIQUES FOR IONOSPHERIC-THERMOSPHERIC INVESTIGATIONS

Edward P. Szuszczewicz
Laboratory for Atmospheric and Space Sciences
Science Applications International Corporation
McLean, Virginia 22102

Abstract

Advances in our understanding of coupled ionospheric and thermospheric responses to energetic inputs are currently limited by an inadequate database. Major problems include the complex interplay of thermospheric winds and electric fields in the Earth's electrodynamic circuit, the lack of routine access to detailed diagnostics of winds, electric fields, electron density distributions and ion composition in the lower ionosphere, and incomplete parameter coverage of cause-effect terms as a function of seasonal, diurnal, solar cycle and geomagnetic controls. "In situ" measurements can provide opportunities for a comprehensive database on the state and condition of the coupled system and the causal mechanisms. These measurements can diagnose the full spectrum of possibilities from the quiescent macroscale domain down to meter- scale irregularity structures known to populate the ionosphere in its most disturbed condition. Aspects of these micro- and macroscale features are reviewed with a perspective intended to develop an understanding of required "in situ" measurement capabilities; and a number of diagnostic techniques are described which form a baseline complement for a rocket or satellite payload. The discussion focuses on traditional techniques which have had broad applications and cover the spectrum of measurements from thermal plasma diagnostics to thermospheric winds and energetic particle spectrometry, with some discussion of limitations, future needs, and the importance of coordinated "in situ" and remote sensing programs.

1. Introduction

While ionospheric research has been actively pursued for more than thirty years, we have yet to understand the interplay of thermospheric and magnetospheric forces in the control of ionospheric plasma distributions and electrodynamic coupling processes. The fundamental limitations in our understanding involve a grossly inadequate database to describe global distributions of electric fields, thermospheric winds, and plasma distributions. We must address these limitations and expand the database to understand ionospheric dynamics carried to the limit of unstable plasma modes and the generation of medium-to-small scale irregularities.

The ionosphere-thermosphere system responds in first order to the major driving forces of solar radiation, magnetospherically-imposed electric fields, and energetic particle precipitation. Some of these particles have direct entry from the solar wind, while others result from prompt or delayed energy transfer processes in the magnetosphere. Second and third order controls of the ionosphere and thermosphere include intricate feedback mechanisms, the streaming-, shear- and current-driven instability processes at high and equatorial latitudes, and the contributions from lower atmospheric disturbances which can propagate to altitudes above 100 km.

To fully understand the ionosphere and the solar, interplanetary, magnetospheric and thermospheric controls means that we must be able to predict the densities and heights of each ionospheric layer at any place and time, and under any condition. We must understand the processes that affect the global electrodynamic circuit through winds, electric fields and plasma layer distributions, and we must be able to predict the dynamics of the ionosphere extended to the limit of unstable geoplasma states and the specification of associated irregularity scale-size distributions.

The sections which follow will briefly review several elements in our current understanding. There will be commentary on causal relationships and ionospheric responses, with a focus on developing an understanding of "in situ" measurement requirements, parameter regimes important to ionospheric-thermospheric investigations, and the demands on spatial and temporal resolution. There will be a review of most frequently used diagnostic techniques and a discussion of future needs.

2. Perspectives on the Physical System and Measurement Requirements

2.1 The F_2 Region

The F_2 region of the ionosphere is generally thought to be laminar and quiescent except under conditions in which unstable plasma modes prevail or there are structured ionization source terms, as in the high-latitude ionosphere. A phenomenological perspective of the various domains is presented in Figure 1, with the illustration broadly intended to depict geomagnetic controls, day/night variability, and coupling to the magnetosphere. The schematic presentation is in the noon-midnight meridian, with the midnight equator at the very left of the figure, the north-polar cap in the center, and the noontime equator to the right. The figure identifies three uniquely different phenomenological domains, the equatorial, midlatitude and polar regions. We move from left-to-right in describing general characteristics of each of the domains.

The nighttime equatorial ionosphere can be made up of structures that span hundreds of kilometers to fractions of a meter [*Szuszczewicz*, 1986; *Singh and Szuszczewicz*, 1984; *Kelley et al.*, 1982]. The disturbed nighttime equatorial condition is defined as

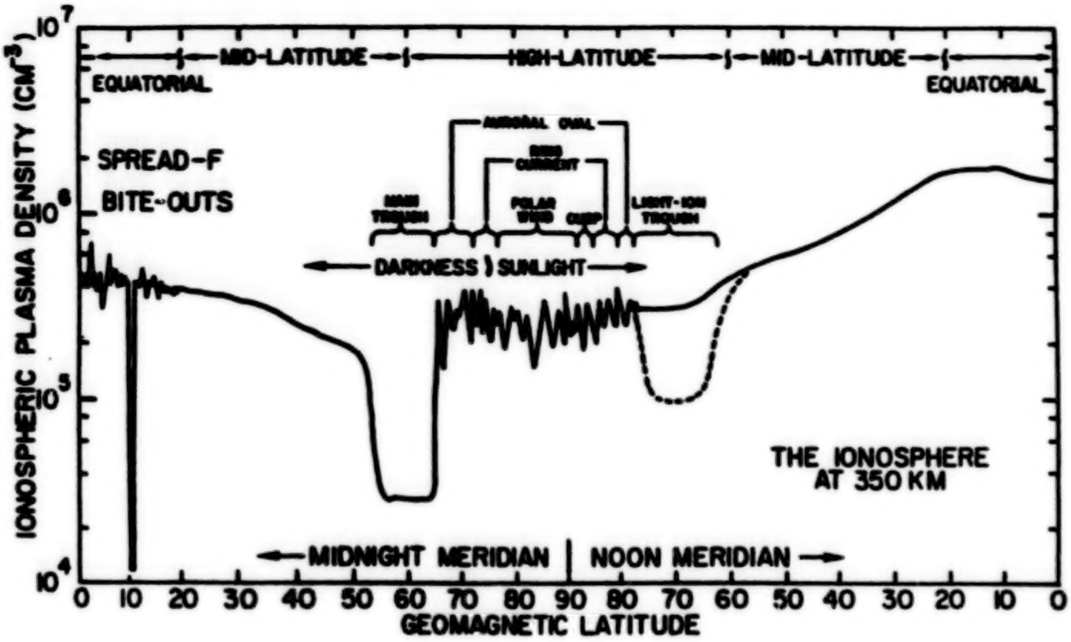


Figure 1. Phenomenological perspective of ionospheric F-region domains portrayed in a noon-midnight meridian [from *Szuszczewicz*, 1986].

equatorial spread-F (ESF) and has come to be characterized by large scale plasma "bite-outs" on the long wavelength end of the spectrum and by meter-scale radar backscatter plumes toward the shortest wavelength portion of the irregularity distribution. The bite-outs are also characterized by dramatic changes in ion composition, with O^+ the dominant ion in the high density region outside the hole, and the molecular ions NO^+ and O_2^+ dominating in lower density region within the holes. The ESF process is a nighttime phenomenon (approximately 2300 ± 300 hours local time), centered about the geomagnetic equator in a belt about ±20° wide in latitude, with seasonal, solar-cycle, and day-to-day variations superimposed.

The generation of irregularities in ESF has been identified with a hierarchy of instability mechanisms (see *e.g.* Figures 2 and 3) generally initiated on the bottomside gradient of the F_2-layer and associated with collisional Rayleigh-Taylor and the $\vec{E} \times \vec{B}$ gradient-drift instabilities (for irregularities in the approximate range 100 m to 2 km). A full accounting of the complete irregularity distribution includes a proposed velocity-sheared Rayleigh-Taylor mechanism at longer wavelengths (1-50 km), universal drift waves at transitional dimensions (appr. 10-200 m) and an assortment of candidate processes (including the kinetic lower hybrid drift instability) in the shortest wavelength regime (< 20 m) [see, for example, *Singh and Szuszczewicz*, 1984; *Kelley et al.*, 1982; *Keskinen et al.*, 1981]. A synthesized perspective of irregularity power spectral density behavior is presented in Figure 3, showing a $k^{-1.5}$ power spectral behavior ($k = 2\pi/\lambda$) at larger wavelengths (2-60 km, the proposed velocity-sheared Rayleigh-Taylor regime), steepening to a $k^{-2.4}$ dependence at intermediate size irregularity structures (100 m to 2 km, the collisional Rayleigh-Taylor spectral region). The transitional wavelength domain (20-200 m) can steepen even further ($k^{-4.8}$ in ionospheric domains with steep density gradients), while the shortest wavelengths ($\lambda \lesssim 20$ m) manifest an almost resonant-like behavior with no systematic power law behavior. Conditions which favor the triggering of ESF include: (i) low plasma density in the field-line-coupled E and F_1 regions, (ii) high altitude for the F_2-peak, (iii) steep bottomside gradients, and (iv) an enhanced $\vec{E} \times \vec{B}$ upward drift velocity.

Irregularity distributions have also been studied in the high-latitude ionosphere where there is special focus on electrodynamics and interactive coupling with the magnetosphere and thermosphere. Unlike the equatorial region, the high-latitude ionosphere can be disturbed on a nearly 24-hour/day basis with inter-altitude coupling effects facilitated by the nearly-vertical geomagnetic fieldline structure. Plasma production at high-latitudes is the result of two sources. The first is the solar-EUV-produced component which varies smoothly across the polar cap, while the second ionization component is identified with energetic electrons which provide the zero-order source term for plasma structure. This is illustrated in the plasma density panel in Figure 4.

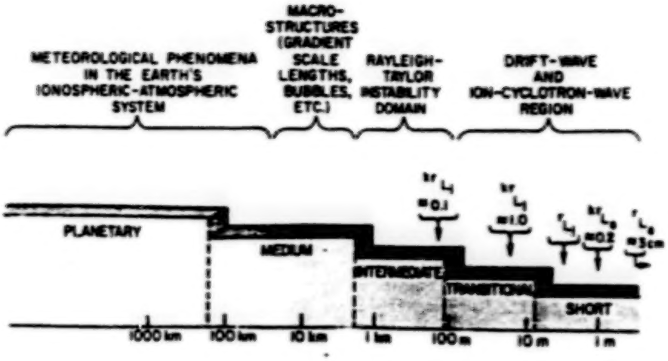

MACRO-STRUCTURES (GRADIENT SCALE LENGTHS, BUBBLES, ETC.)
METEOROLOGICAL PHENOMENA IN THE EARTH'S IONOSPHERIC-ATMOSPHERIC SYSTEM
RAYLEIGH-TAYLOR INSTABILITY DOMAIN
DRIFT-WAVE AND ION-CYCLOTRON-WAVE REGION
$kr_{L_i} = 0.1$
$kr_{L_i} = 1.0$
r_{L_i}
$kr_{L_e} = 0.2$
$r_{L_e} = 3$cm
PLANETARY
MEDIUM
INTERMEDIATE
TRANSITIONAL
SHORT
1000 km
100 km
10 km
1 km
100 m
10 m
1 m

Figure 2. A simplified synoptic perspective of the full domain of equatorial F-region irregularities with potential mechanisms. Characteristic lengths are the wave number $k = 2\pi/\lambda$, and the ion and electron gyroradii r_i and r_e for $M_i = 16, T_e = 1350°K, T_i \sim 500°K$, and $B = 0.3G$ [from *Singh and Szuszczewicz*, 1984].

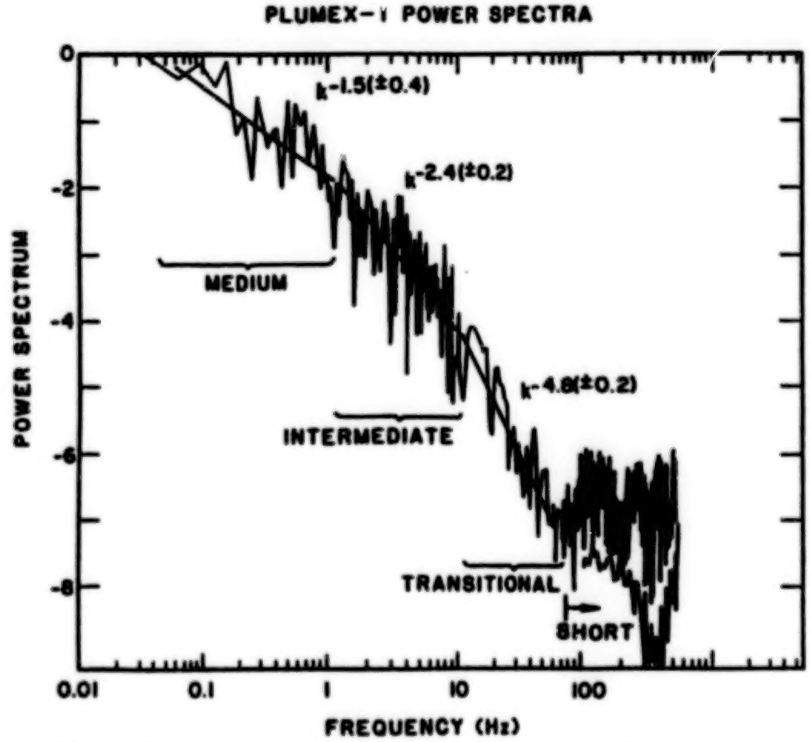


Figure 3. Power spectral distributions of equatorial spread-F irregularities from 50 km scale sizes to dimensions near 1 m. Abscissa is rocket-borne telemetry sampling rate which has a different scale-size conversion for the synthesized segments of the power spectrum. 0.1 Hz and 100 Hz correspond approximately to 20 km and 20 m, respectively [from *Singh and Szuszczewicz*, 1984].

THE HIGH-LATITUDE IONOSPHERE PHENOMENOLOGY, IRREGULARITY DISTRIBUTIONS, TRANSPORT AND MAGNETOSPHERIC COUPLING

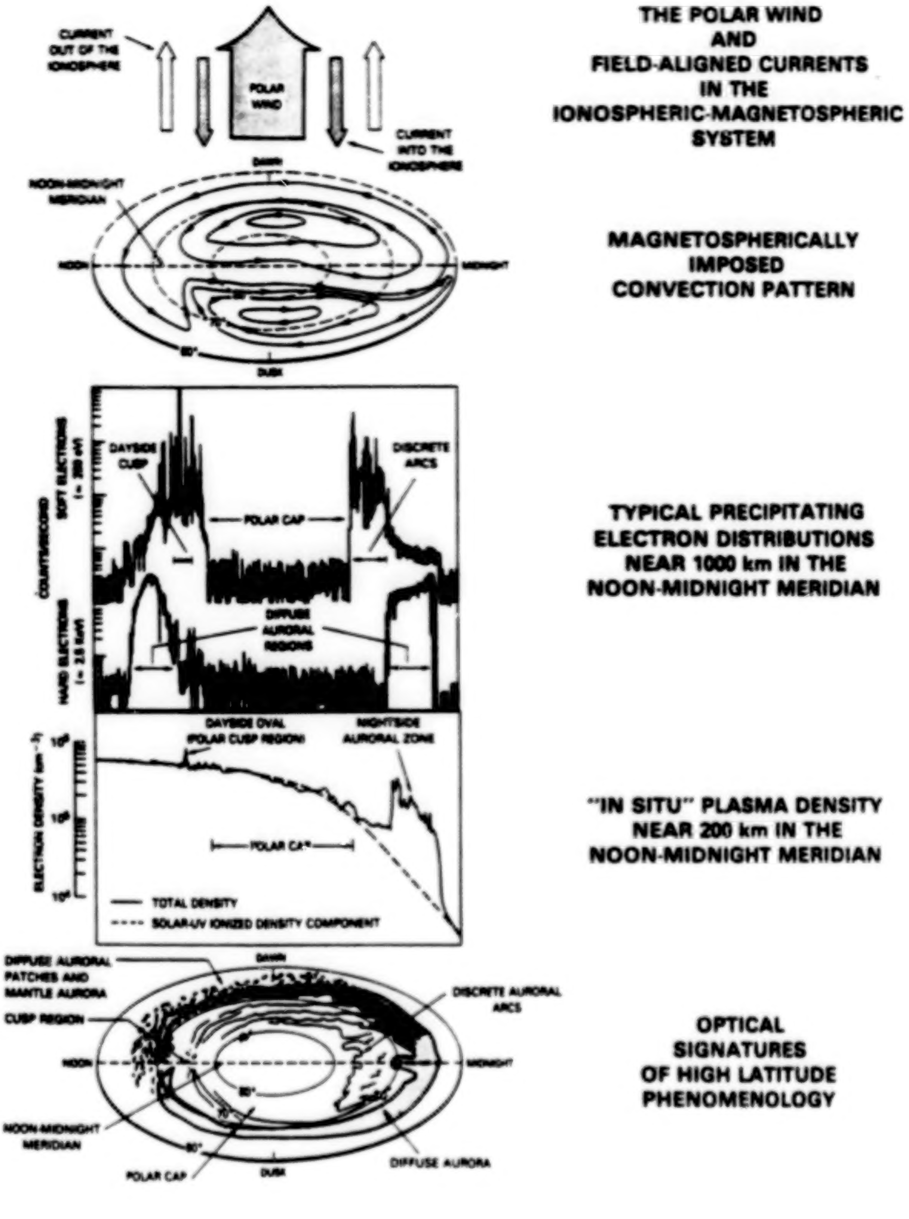


Figure 4. Composite illustration of high latitude ionospheric processes including optically-derived phenomenology, plasma density structures, precipitating particle distributions and coupling to the magnetosphere through a convective electric field, field-aligned currents and the polar wind. The format is in coordinates of magnetic local time and magnetic latitude with specific adaptations from the works of *Akasofu* [1981], *Rodriguez and Szuszczewicz* [1984], and *C. I. Meng* (private communication, 1983).

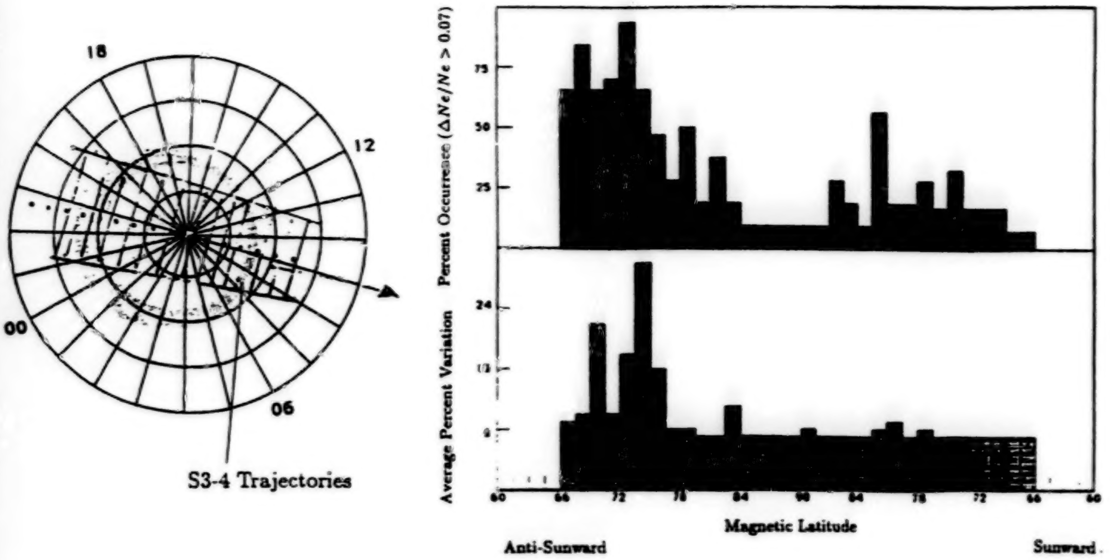
The electron precipitation at high-latitudes is variable in its energy and density distributions, horizontal extent and temporal characteristics. As a result, it creates horizontal and vertical ionospheric plasma structure by virtue of its own spatial, temporal, and energy distributions, and it provides a free energy source for a number of current-driven and two-stream ionospheric plasma instability processes. Once created, local plasma structures are subjected to magnetic-field controls, a magnetospherically-imposed convection electric field and interactions with superimposed winds and current systems (see top panels of Figure 4).

The plasma structure produced in the high-latitude ionosphere ranges from hundreds of kilometers to meters [*Rodriguez and Szuszczewicz*, 1984; *Curtis et al., 1982*; *Huba and Ossakow*, 1980; *Keskinen and Ossakow,* 1982; *Phelps and Sagalyn*, 1976; *Rino et al.*, 1978; *Singh et al.*, 1985 and 1987; *Szuszczewicz et al.*, 1982] but unlike the situation at equatorial latitudes, concerted efforts have yet to fully characterize the spectral behavior of irregularity distributions in each of the high-latitude regions. One should expect that irregularity distributions will be different in the dawn-, dusk-, noon-, and midnight ovals, and that there will be an altitude-dependence in the spectral distributions. Some of these behaviors have been seen in S3-4 satellite "in situ" plasma density observations across the summer North Pole and winter South Pole domains (see *Szuszczewicz et al.*, 1982, for orbital details and measurement characteristics). Sample data from *Singh et al.*, 1987, are presented in Figures 5 and 6 and their review is pertinent to irregularity scale size distributions, associated global-scale transport properties, and ion-chemical controls.

North Polar ionospheric observations summarized in Figures 5 and 6 were in the molecular-ion (*i.e.*, NO^+ and O_2^+) dominated altitude region below 200 km, while the South Polar ionospheric region was sampled at somewhat higher altitudes (approximately 260 km) where O^+ tended to be the dominant constituent. These differences had a major impact on the resulting observations. Lower altitude irregularities ($\delta[NO^+]$) are shorter lived and therefore show effects of local processes; higher altitude irregularities ($\delta[O^+]$) are longer lived and can be transported large distances from their source domain. This is evident in the percent occurrence plots of Figure 5. Low altitude, $\delta[NO^+]$ irregularities are dominant in the day- and nightside oval but not in the cap, a direct result of particle precipitation morphologies. On the other hand, longer lived high altitude ($\delta[O^+]$) irregularities are transported throughout the South Polar cap and the resultant percent-occurrence plot shows no distinguishing features for any of the domains.

The differences in irregularity lifetimes are also manifested in the irregularity scale-size distributions. To demonstrate this, we focus on a synoptic perspective of the power

NORTH POLE



SOUTH POLE

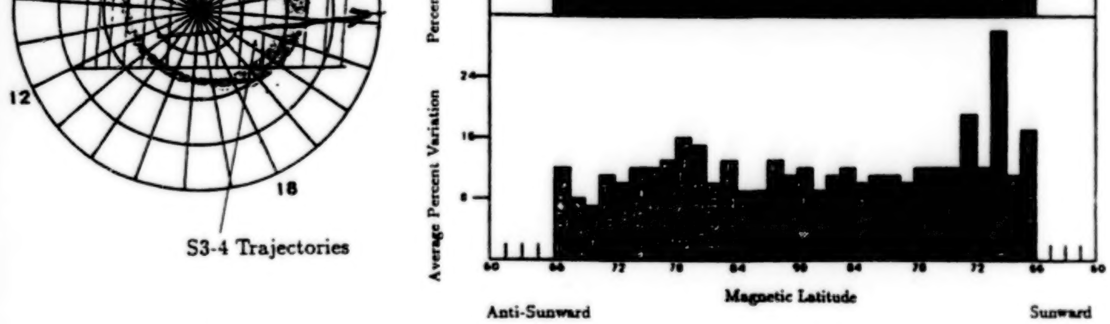


Figure 5. (Left) The S3-4 satellite trajectories used in the data analysis, with modelled oval (hatched) in MLAT-MLT coordinates. The black histogram panels show percent occurrence of density variations greater than 7%, and the gray histogram panels show averaged percent-variations of electron densities. Both histograms show the dependence on magnetic latitude [from *Singh et al.* (1987)].

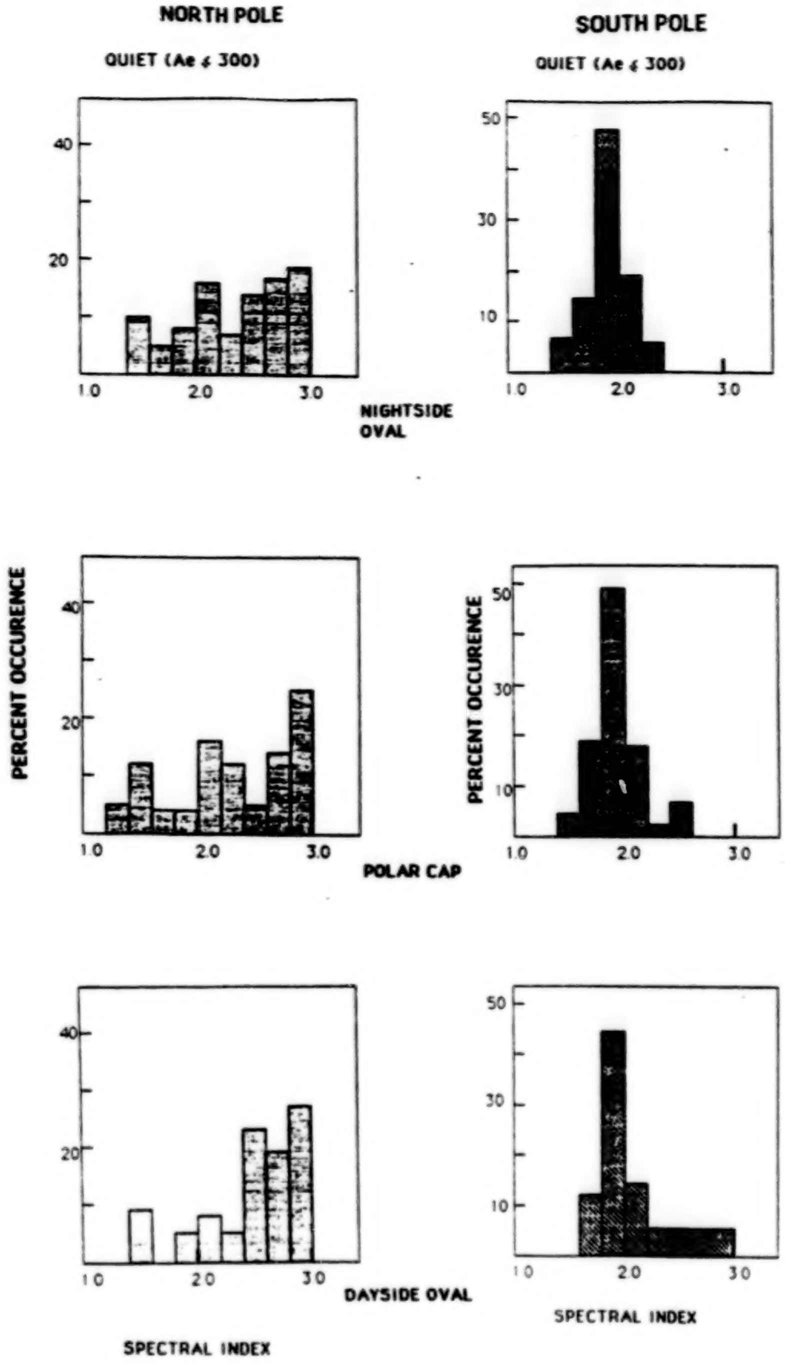


Figure 6. Percent occurrence histograms showing spectral indices for high-latitude irregularities in the nightside/dayside oval and in the polar cap under geomagnetically quiet conditions (AE<300) [adapted from *Singh et al.* (1987)].

spectral distributions of the observed irregularities ($P(k) \propto k^{-x} = (2\pi/\lambda)^{-x}$) in the 1-50 km wavelength regime (λ = irregularity wavelength) as published by ***Singh et al.*** [1987]. A typical result is presented in Figure 6, which includes histogram plots of measured spectral indices ("x" in $k^{-x} = (2\pi/\lambda)^{-x}$) in the polar cap ionosphere. The left panels of Figure 6 summarize the lower-altitude sunlit North Polar ionospheric results, while the right panel of Figure 6 summarizes the higher-altitude South Polar counterpart. The abscissa is the computer-fitted spectral index, while the ordinate is the observed frequency of occurrence. The differences across the polar regions are seen to be rather dramatic. There is a well-defined most probable spectral index at $x = 1.9$ in the higher-altitude South Polar region, while the lower-altitude North Polar data display a broad range of indices with no clearly-defined most-probable behavior. The differences can be identified with "ordered" and "unordered" plasma processes. At higher-altitudes (the data in the South Polar region), the irregularities are much longer lived and more likely to manifest susceptibility to well-defined plasma instability processes with predictable irregularity spectral distributions. On the other hand, in lower altitude regions (the data in the North Polar regions) where irregularity distributions are primarily a function of the spatial distributions of precipitating particles (largely a nondeterministic process), the spectral indices will tend to be broad ranging, without an easily-identified most-probable value. More detailed comparisons are made for each of the high-latitude domains, including the dayside and nightside ovals, in the work of ***Singh et al.*** [1985 and 1987].

Results like those in Figures 5 and 6 provide a database for understanding the causal mechanisms responsible for the full spectral distribution of irregularities and their global scale transport properties in the high-latitude ionosphere. But because of the near-vertical magnetic field, the complex nature of energetic-particle precipitation patterns, and the myriad of inter-altitude coupling mechanisms, substantially more work must be done, including vertical profile information from rocket-borne sensors and ground-based diagnostics. While ground-based diagnostics are site-specific, they are very useful for long-term systematic studies [*e.g.*, ***Baron et al.***, 1983]. When coupled with similar diagnostic capabilities at other locations, they can also contribute to the understanding of large-scale ionospheric coupling processes and the role of longitudinal and latitudinal transport in the polar ionosphere.

2.2 The E- and F_1-Regions

The importance of winds and electric fields has long been recognized because of their roles in controlling ionospheric plasma density distributions and E- and F-region dynamics. The equatorial and low-latitude ionosphere is dominated by electric fields generated by atmospheric tidal motion in the E-region, a process referred to as the "E-

region dynamo" [see *e.g. Akasofu and Chapman*, 1982; *Hines et al.*, 1974]. While the source for these fields is in the 100-150 km region at high-middle to low-latitudes, the fields map equatorward to higher altitudes along geomagnetic fluxtubes and produce $\vec{E} \times \vec{B}$ drifts in the ionospheric plasma. These $\vec{E} \times \vec{B}$ drifts have a dominant diurnal component, reflecting a solar-diurnal tide in the E-region driven by solar heating. The dynamo system also supports horizontal currents that have latitudinal, longitudinal and seasonal variations. These current systems in first order depend on the ionospheric plasma densities and electric fields (*i.e.*, $j = \sigma E \propto neE$).

There are several serious difficulties in our understanding and modelling of the dynamo fields and current systems. First, the plasma density profiles are seldom (if ever) as clean and simple as our traditional descriptions of laminar and quiescent E and F_1 layers. This is particularly true under dynamic, unstable or otherwise irregular conditions and is most dramatically manifested at night. It can also be true under what would otherwise be considered a nominally-quiescent condition. There is also the problem involving the "anisotropic" nature of conductivities, as represented by the longitudinal σ_0, and transverse Pedersen σ_1, and Hall σ_2 contributions. Each conductivity varies differently with the altitude distributions of the ionospheric plasma, the magnetic field strength, ion composition, particle temperatures and the neutral atmospheric densities. The situation is further complicated by the fact that the electric field consists of several distinct components, such as the corotation field, the superimposed magnetospheric fields, the wind-driven dynamo fields ($\vec{E}_d = -\vec{V} \times \vec{B}$) and the polarization fields E_p. The result is a net field and a current system that depend critically on the height distributions of N_e, as reflected in the linear dependence of conductivities σ_0, σ_1 and σ_2 on the local ionospheric plasma densities [see *e.g., Akasofu and Chapman*, 1982].

Perhaps the most important conclusion to be drawn from what has been said is the globally-significant role played by winds and ionospheric plasma distributions in the 100 - 200 km altitude region. This is the source region for a dynamo-generated potential of $(10)^4$ volts with a current of $(10)^5$ amperes [*Banks et al.*, 1988]. Because the fields are carried upwards along geomagnetic fluxtubes, the winds and plasma distributions in the dynamo region can influence the dynamics at higher altitudes and at locations horizontally removed from the source region by 100's to 1000's of kilometers. The electric field coupling mechanisms operate on the remote ionospheric plasma which in turn couples to the thermosphere through ion-neutral collisions. Attempts to develop global models of these coupling processes are hampered by the fact that the 100 - 200 km regime is the least explored geoplasma domain [*Banks et al.*, 1988; *Scarf et al.*, 1988]. This region has been studied primarily by site-specific ionosonde, radar and rocket investigations, since it is generally not accessible to long-lived satellite investigations which can provide large synoptic data profiles. The overall result is a lack of appropriate data to study this key region in the Earth's global electrodynamic circuit.

The ionospheric-thermospheric community must address the lack of data and associated lack of understanding of processes that affect the global electrodynamic circuit through winds and plasma layer distributions in the 100 - 200 km regime. There should be a focus on the development of empirical specifications of thermospheric winds and plasma layer distributions in that region, with a clear recognition that those distributions are rarely (if ever) laminar and quiescent. A simple example in support of this last statement is sporadic-E (often designated simply as Es), which has long been a recognized component of plasma distributions in the lower ionosphere. Originally thought to be a mid-latitude phenomenon driven by wind-shear convergence of long-lived metallic ions [*Whitehead*, 1970; *Matsushita and Smith*, 1972 and 1975; *Strobel*, 1974], its effect on the ionospheric height profile was considered to be largely confined to altitudes near 105 km, with Es layer-thicknesses of about 1 - 5 km and peak densities at times as great as those at the F_1 and F_2 peaks (*i.e.*, of the order of $(10)^5$ cm^{-3}). This early idea of layer height, thickness and intensity is conveniently represented in Figure 7 by the rocket data of *Szuszczewicz and Holmes* [1977] collected in the mid-latitude ionosphere above White Sands, N.M. While the rocket data presents a clear example of the Es layer parameters (providing $N_e, \delta N_e, T_e, M_i$, etc.), attempts to develop global perspectives have relied almost entirely on ionosonde measurements of critical frequencies (foEs [Hz] $= 8.9(10)^3\sqrt{NmEs\ [\text{cm}^{-3}]}$) and virtual heights ($h'$ Es $\approx$ hmEs near 105 km). The early studies by *E. Smith* [1957, 1961 and 1976] and *Leighton et al.* [1961] are the only efforts to date which develop such a perspective. Those studies provided maps of Es occurrence probabilities and height distributions based on monthly mean ionosonde observations. That early work and continuing rocket investigations [*Smith*, 1970; *Szuszczewicz and Takacs*, 1979] began to make it clear that the developing concept of sporadic-E was not limited to mid-latitudes but instead had a wide latitudinal extent including equatorial and high-latitude domains. Furthermore, accumulated data has made it clear that Es is not a single layer phenomena, but a process involving multiple layers of varying thicknesses, dependent upon wind-shear convergence efficiencies that extend from regions very close to the magnetic equator up to high-mid-latitudes. The data suggest that layers at or near 105 km could be part of a deterministic process involving wind shear convergence in a global system of diurnal and semi-diurnal tides. This idea is best illustrated in the work of *Shen et al.* [1976] and *Mathews and Bekeny* [1979] at the Arecibo Observatory. Figure 8 presents some of the *Shen et al.* results, where it is clear that there is a time-dependent layer-formation and transport process initiated on the bottom-side of the F_2 region. That process, believed to be driven by wind shear convergence of ambient ions, begins near 180 km with a phase

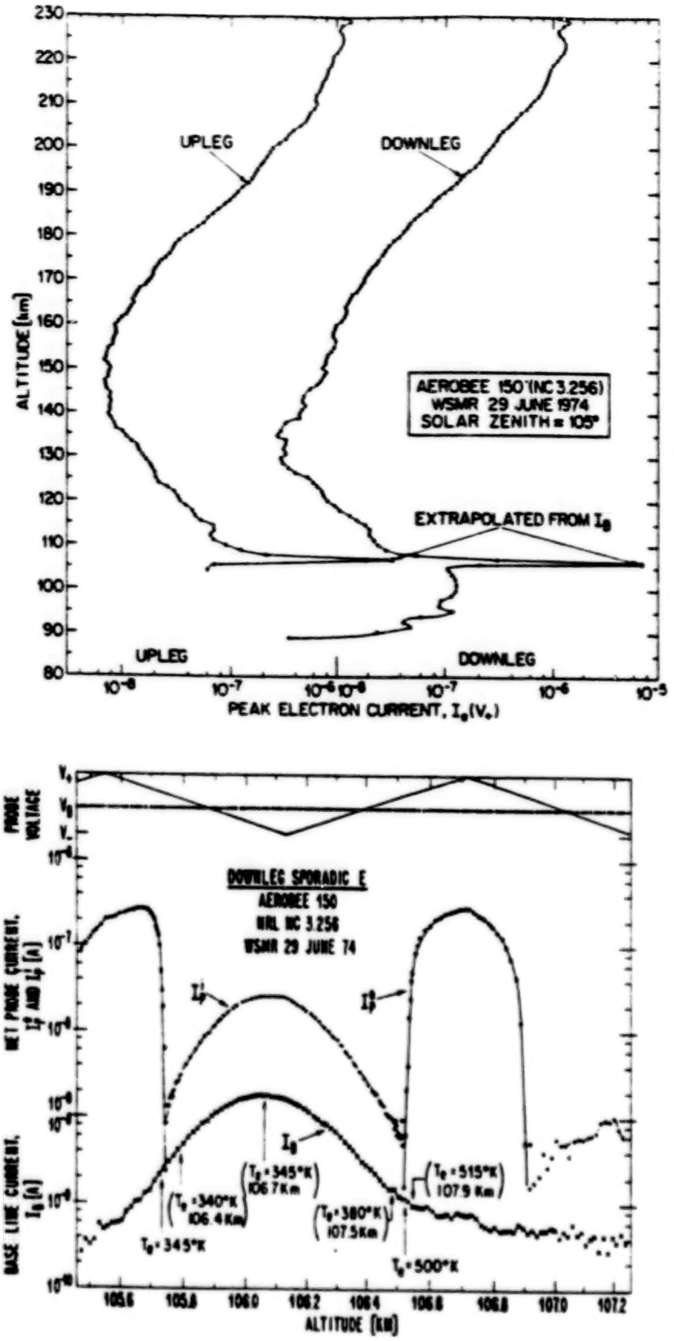


Figure 7. Electron temperature and plasma density data from a sounding rocket flight through a sporadic-E layer at WSMR, NM; top panel shows upleg and downleg total probe current ($I_e(V_+) = 10^{-5}$amp $\Rightarrow N_e \times 3(10^5)$cm^{-3}); bottom panel shows an expanded view of ion, electron and total probe current along with measured electron temperatures [from *Szuszczewicz and Holmes*, 1977].

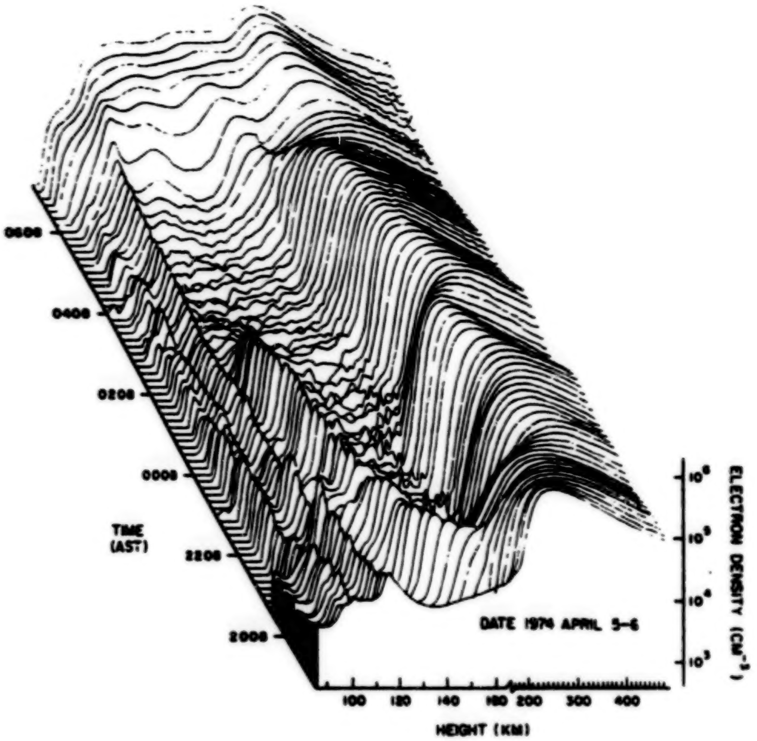


Figure 8. Arecibo incoherent scatter radar density profiles showing nighttime intermediate layer formation and descent [from *Shen et al.*, 1976].

coherent downward transport of the "collected" plasma, resulting in the formation of quasi-stationary layers in the 100 - 120 km altitude regime. The "converged" high altitude layers and their associated downward transport have given rise to the names "descending," "intermediate" and "sequential" layers; and it appears that they play an important role in the formation of "classical" low-altitude Es-layers through a process that is not "sporadic," but deterministic [***Matthews and Bekeny***, 1979]. This idea is compatable with the developing thermospheric work of R. Roble [private communication, 1989], wherein his global scale models are in concert with descending horizontal wind-shear nodes in the 100 - 200 km region. We present a selection of his results in Figure 9.

The interplay of electric and thermospheric forces in the formation and control of these ionospheric layers is both subtle and intricate. According to wind shear theory, winds alone are not the only controlling force, but instead may play an interchange role with electric fields. This is simply represented in the treatment of *Lanchester et al.* [1989] who show that the vertical ion velocity V_Z, the thermospheric wind $\vec{U}$, and local electric fields $\vec{E}$ are interrelated by the expression:

$$\left(1+\frac{\nu^2}{\omega^2}\right)V_z = \left[\frac{1}{B}\left(E_E+\frac{\nu}{\omega}E_{\perp N}\right)+\left(\frac{\nu}{\omega}U_E+U_S\sin\mathrm{I}\right)\right]\cos I+\left(\frac{\nu^2}{\omega^2}+\sin^2 I\right)U_z.$$

Accordingly, layers form at convergent nulls of the vertical ion drift V_z; *i.e.*, where $V_z = 0$ ($V_z < 0$ above and $V_z > 0$ below). Whether or not $V_z = 0$, and a convergent null is formed, depends on the horizontal neutral wind components U_E and U_S, and the electric field components E_E and $E_{\perp N}$. The subscripts E, S and $\perp N$ define directions magnetically eastward, magnetically southward, and northward and perpendicular to $\vec{B}$, respectively. Contributing terms include the ion-neutral collision frequency ν, the ion gyrofrequency ω and the dip angle I. The relative importance of the terms is readily seen to be strongly height dependent, through ν and ω. Generally speaking, $\nu < \omega$ ($\nu > \omega$) at heights above (below) 125 km.

The purpose in drawing attention to the above equation is to illustrate some of the intrinsic and subtle problems in trying to measure and model layer formation and dynamics on a global scale. We have the interplay of the wind and electric field components, complicated by a non-existent and/or impoverished database on global distributions of $\vec{E}$ and $\vec{U}$ in the altitude range in question. This problem even has serious aspects at F-region heights, where the traditional global-scale electric field model of *Richmond et al.* [1980] has been found grossly inadequate [*Szuszczewicz et al.*, 1988; ***Schunk and Szuszczewicz***, 1988]. The modelling problem is further exacerbated by issues involving ion composition. Conventional wisdom has always required long-lived metallic ions to maintain the "sporadic" layers at high densities and at low altitudes.

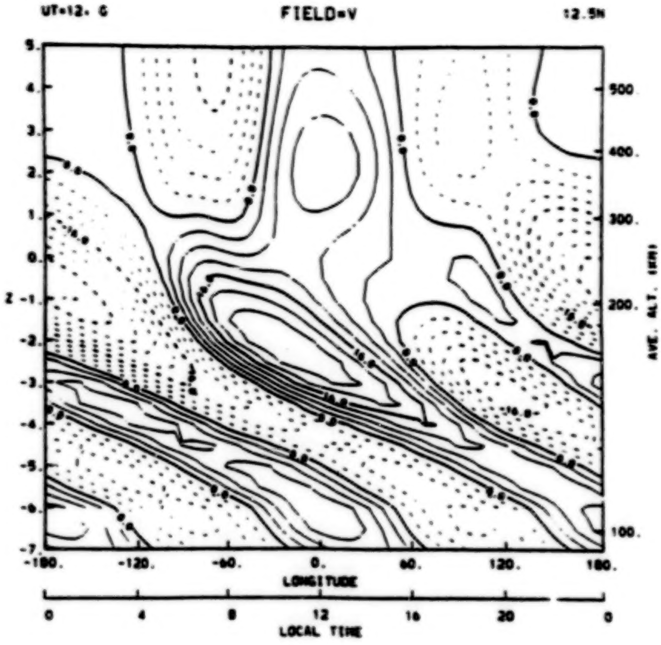


Figure 9. Meridional neutral winds from R. Roble's global thermospheric circulation model [*R. Roble*, private communication 1989]; solid contours represent northward winds, dashed contours represent southward winds at UT=12 and at 12.5° NLat.

But at F_1 and lower-F_2 heights the ion populations are dominated by O^+ and varying combinations of moleculars NO^+, O_2^+, N_2^+, NO^+, etc. Convergence and "pile-up" of moleculars would ultimately lead to their loss through increased rates of dissociative recombination; and descending layers of O^+ would see their ultimate demise through the rate-limiting reaction $O^+ + N_2 \rightarrow NO^+ + N$, and subsequent dissociative recombination of the NO^+ product. As the "intermediate" layers descend, their continued existence at high densities demands a steady accumulation of metallics while losing O^+ and the molecular ions to electron recombinations [*Mathews and Bekeny*, 1979].

The question of the overall availability of metallic ions has yet to be resolved. One might expect that semi-diurnal tidal motions might sweep the ionosphere clean of metallic ions twice a day, but *Lanchester et al.*, [1989] and *Whitehead*, [1970] suggest that meteoric input is insufficient to maintain a regular daily cycle of descending layer phenomena. There is evidence, however, that lack of phase coherence in the contributing tidal modes at altitudes near 125 km can break up the layers and redistribute the metallic ions. There is also the suggestion [*Mathews*, 1988] that the metallic ions which are swept up and deposited at altitudes below 100 km are transported equatorwards and "recycled" to the ionospheric F-region domain through the fountain effect (a well-documented and accepted process for upward transport of metallic ions at and near the equator [*Hanson et al.*, 1972; *Grebowsky and Reese*, 1989] resulting in a redistribution to higher north/south latitudes). Ultimately, however, the total metallic-ion content in the 100-200 km domain traces its origin to meteor ablations and seasonally-dependent meteor showers [*McKinley*, 1961].

These complexities make clear the current intractability of a meaningful understanding of the global distributions of the ionospheric plasma in the 100 -200 km region. There is a need for an intense measurement program to specify the forces and responses in this region of the Earth's electrodynamic circuit. Such a measurement program could provide a definitive test of the concept that "dynamic," "irregular" or otherwise "sporadic" E- and F_1 processes are part of a systematic chain of forces that involves various tidal modes of thermospheric winds (perhaps even a quarterdiurnal component [*Harper*, 1977; *Tong et al.*, 1988]), dynamo driven electric fields and ion chemistry. With this expectation, a detailed and coordinated measurement activity could hold the key to a significantly improved understanding of the ionospheric dynamo region and the ultimate treatment of the entire ionospheric domain as one dynamic and chemically-active fluid.

3. Baseline Techniques for "In Situ" Particle and Field Measurements

3.1 An Overview on Requirements

The previous sections have attempted to establish a geophysical framework for required "in situ" measurements in the ionospheric-thermospheric system. The focus was on dynamic, irregular conditions since they stress the spatial, temporal and parameter-range capabilities of any given diagnostic device. In general, if a measurement technique can meet the requirements of a dynamic environment, it can also provide the baseline data to define the zero-order, quiescent condition. This is not to underplay the importance of quiescent or moderately-dynamic states, for they define the very modes which can precipitate unstable plasma conditions and a broad range of irregularity distributions. Indeed, an accurate baseline specification of the global distribution of "quiescent" electric fields and thermospheric winds is non-existent.

The general guideline for measurement techniques therefore tends to be relatively straightforward. The parameter set must be complete (*i.e.*, we must measure all cause and effect terms, where possible); the dynamic range and accuracy must match the needs in describing the physical system; and the spatial and temporal resolution must span the range of possibilities from quiescent to fully turbulent. In general, we should measure the primary energy inputs involving solar EUV-UV radiation and the fluxes of precipitating particles at high latitudes. We must be able to measure the plasma density from 10^3 to $4(10)^6$ cm^{-3}, the electron temperature from 300-4000°K (and allow for non-Maxwellian, bi-Maxwellian or otherwise anomalous energy distributions), the electric fields from dc to approximately 50 Mhz and irregularity distributions from ten kilometer scale sizes to the meter domain. Ion composition must be resolvable from 1 - 64 amu, and ion energy determination must include the capability to detect anisotropic distributions as they can be manifested in the formation of ion conics. Ion composition measurements must also have spatial and temporal resolutions which match the prevailing conditions at high latitudes, within and across intermediate layers, and under conditions of equatorial spread-F. There are also high demands placed on the measurement of thermospheric winds. The dynamic range must run from 0 - 1000 m/sec with an accuracy approaching ± 15 m/sec and a 1 km spatial resolution over the 95-600 km altitude domain.

In what follows, we describe the "traditional" set of diagnostic techiques that have been used in accumulating the bulk of data available to date. In several circumstances we will address the issue of resolution and summarize some of the more important development requirements in the final section of this paper. Our treatment here divides the measurement techniques into three categories: (1) those which measure the state and condition of the ionosphere, (2) those which measure the state and condition of the thermosphere, and (3) those which define energetic inputs, field controls, transport and geophysically-significant boundaries.

3.2 Measurement Techniques for the State and Condition of the Ionosphere

Langmuir-type probes. The ionospheric manifestations of ion-neutral coupling, dynamic fields, and magnetospheric inputs are embodied in the global distributions of the ion/electron densities and their associated energies. The complete specification includes not just the densities and energies, but the details of ion composition, gradient scale lengths and irregularity distributions.

In determining the state and condition of the ionosphere, the technique which appears to have received the broadest application is that of the electrostatic probe, better known as the Langmuir probe because of the pioneering work of Irving Langmuir in the first quarter of this century [***Mott-Smith and Langmuir***, 1926; ***Chen***, 1965]. Langmuir probes on Explorer 22, 32 and Aloutte-II [***Brace***, 1969; ***Brace et al.***, 1967;] as well as on the AE-C and DE-2 spacecraft [***Brace and Theis***, 1981; Kozura et al., 1986; ***Curtis et al.***, 1985; ***Fontheim et al.***, 1987] have provided much of the information available today on electron density and temperature morphologies in the global-scale ionosphere. The DE-2 Langmuir probe data also provided opportunities to study density irregularites in the wavelength interval covering 30 - 170 km in a morphological investigation that considered equatorial spread-F and high-latitude phenomenologies [***Hoegy et al.***, 1984]. A specialized Langmuir probe (referred to as a Pulsed Plasma Probe, P^3) on the S3-4 satellite complemented the DE-2 irregularity study by extending the investigation to a broader range of scale sizes, covering the full domain from 100's of km to 20 meters [***Rodriguez and Szuszczewicz***, 1984; ***Singh and Szuszczewicz***, 1984; ***Singh et al.***, 1987; ***Szuszczewicz et al.***, 1981 and 1982].

The Langmuir probe is most simply described as a conductor (generally of planar, cylindrical, or spherical geometry) which collects current from a plasma when a voltage is applied. The current drawn from the plasma by the probe is a function of the probe size and geometry, the probe voltage, and the plasma properties of charged-particle number densities, particle distribution functions, and collision frequencies. Consequently, a current-voltage characteristic of a probe imbedded within a plasma is potentially rich with information about that plasma. If one understands the behavior of plasmas in the presence of an electrostatic probe, then in principle the plasma parameters mentioned above can be extracted from a probe characteristic.

The Langmuir probe is the smaller electrode of a two-electrode configuration (*i.e.* the probe and its reference electrode forming a closed circuit with the plasma) with the ratio of the two areas approaching a value which for all practical purposes should be considered infinite (see insert in Figure 10). In a laboratory situation the reference electrode can, in fact, be the container of the plasma volume, while in a spaceborne application the rocket or satellite body functions as the reference electrode. A typical spaceborne cylindrical Langmuir probe is 20 cm in length and 1 mm in diameter. When the two electrodes are in contact with a plasma, a current will pass between them which is a function of an applied voltage difference and the prevailing plasma conditions. When this current is plotted as a function of the applied voltage, the resulting curve is referred to as the probe characteristic. Figure 10 shows a schematic representation of a Langmuir-probe circuit as well as a typical I-V characteristic. In conventional analysis procedures [*Chen*, 1965] N_e is determined from the electron-saturation portion of the characteristic ($V_a >$ plasma potential), T_e is determined from the transition (retarding-field) region (floating potential $\lesssim V_a <$ plasma potential), and mean-ion-mass is determined from the ion-saturation portion of the curve ($V_a <$ floating potential).

Conventional Langmuir probes (like those on the Explorer, AE and DE series of satellites) typically employ a continuous time-dependent voltage function (*e.g.* sine wave, ramp, sawtooth, etc. as in Figure 11A) to generate current-voltage (I-V) characteristics from which N_e, T_e, etc. are extracted. The time-honored conventional approach has broad applicability in fully-Maxwellian plasmas with a single characteristic electron temperature T_e and ion temperature T_i, and with temporal variability on time scales much larger than the period required to generate a single I-V characteristic. When these conditions are violated, more sophisticated approaches to Langmuir probe diagnostics are required. One such approach is that of the Pulsed Plasma Probe, P^3 [*Szuszczewicz and Holmes*, 1975; *Holmes and Szuszczewicz*, 1981]. The P^3 approach employs a series of voltage pulses ($\sim$ 100 μs wide) which follow a sawtooth envelope (see Figure 11B). During the interpulse period ($\sim$ 900 μs), the probe is held at a fixed baseline voltage level V_B which is generally positioned in the ion or electron saturation region of the probe's I-V characteristic. The long (relative to the 100 μs pulse) baseline voltage level not only stabilizes surface conditions between pulses (eliminating hysteresis effects and erroneously high measurements of electron temperature) but running measurements of baseline currents provide a measure of plasma density variations and the fundamental data for the determination of density fluctuation power spectra and associated turbulence structure [*e.g., Rodriguez and Szuszczewicz*, 1989; *Singh and Szuszczewicz*, 1984]. The continuous measurement of baseline current also allows the unfolding of a Langmuir I-V characteristic when density variations occur on a time scale short compared to the probe's sweep period [*Szuszczewicz et al.*, 1979; *Szuszczewicz and Holmes*, 1977]. Such conditions would prevail as an instrumented satellite passed through a narrow auroral arc, or a rocket trajectory carried an instru-

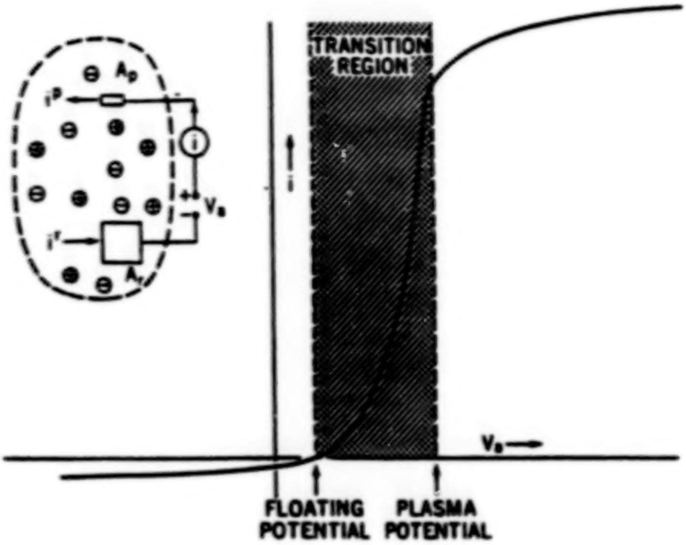


Figure 10. Schematic representation of a Langmuir probe circuit and a corresponding current-voltage characteristic [from *Szuszczewicz*, 1972].

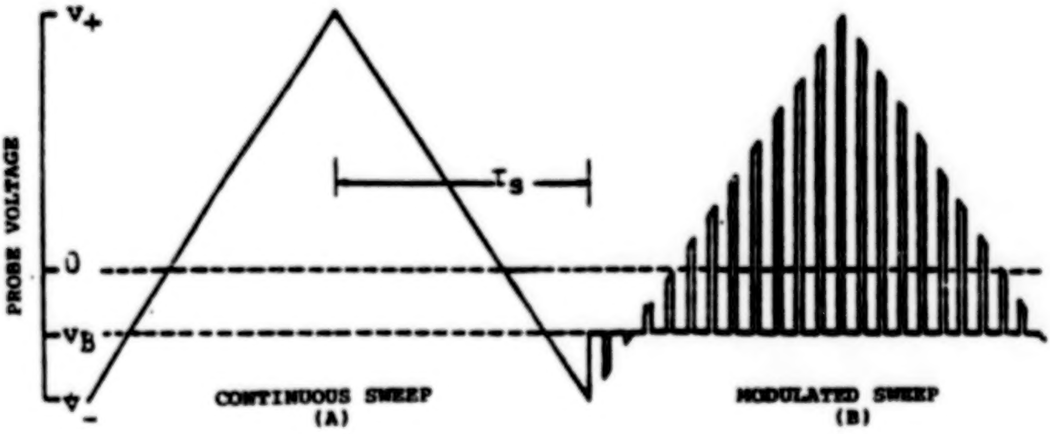


Figure 11. Continuous (panel A, the conventional Langmuir probe) and pulsed-modulated (panel B, the P^3 format) sweeps. [See *Szuszczewicz et al.*, 1982.]

mented payload through a series of narrow intermediate layers [see *e.g.* Figure 7 and ***Szuszczewicz and Holmes*** (1977) for P^3 measurements of T_e across narrow E_s layers].

P^3 is also unique in the measurement of multi-component electron energy distributions. This is an important issue, since it is often the experience of experimental laboratory and space plasma physics [*e.g.*, ***Szuszczewicz***, 1983] that turbulence is associated with anomalous electron energies, including bi-Maxwellian and "bump-on-tail" distributions. When irregularities exist, a conventional Langmuir I-V characteristic could not differentiate between anomalies in density variations and energy distributions. The P^3 was designed to do this. Using the baseline current levels for tracking density fluctuations in a turbulent beam-plasma interaction region, P^3 diagnosed a suprathermal electron component [***Szuszczewicz** et al.*, 1983] superimposed on a cold Maxwellian in a two-stream interaction process very much like that which can occur in the high-latitude ionosphere.

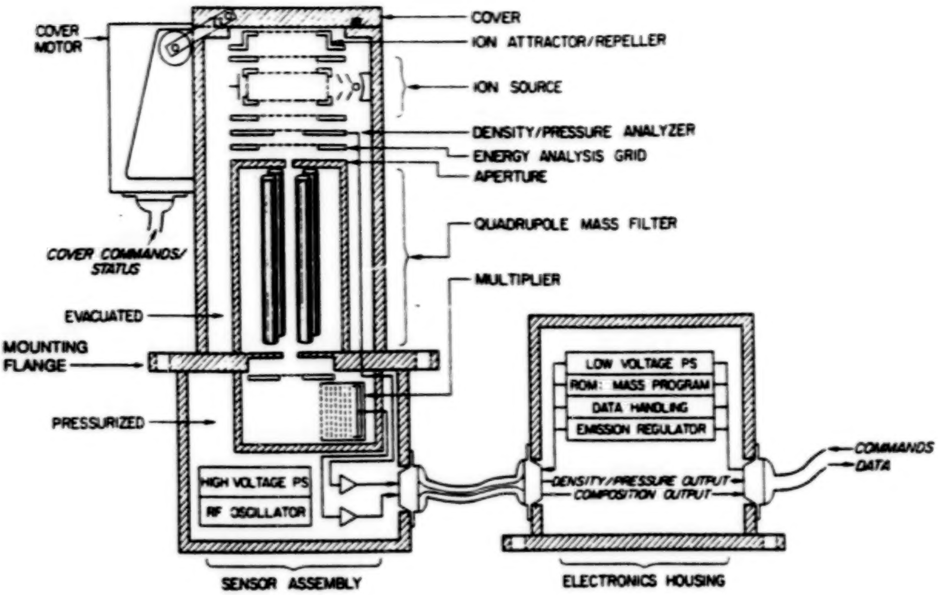
If two such probes are used, their baseline voltages may be chosen to lie in the ion and electron saturation regimes, respectively. The ratio of the two baseline currents will allow rapid tracking of mean-ion-mass variations, their associated power spectral distributions and effects on instability processes [***Szuszczewicz** et al.*, 1982]. This capability is expected to be of particular interest across auroral boundaries, polar cap "blobs," the mid-latitude trough, and ESF "bubbles." Useful applications also include rocket-borne diagnostics of $N_e, \delta N_e, T_e$ and δM_i in and across intermediate layers in the E and F_1-regions of the ionosphere.

Ion (and Neutral) Mass Spectrometers. While Langmuir-type probes define the state and condition of the electron component of the ionospheric plasma and its mean-ion-mass, mass spectrometers are necessary to determine the abundance of each ion species as a function of position along a satellite or rocket trajectory. Unlike the Langmuir probe sensor, which is a simple cylindrical, spherical or planar electrode suitably deployed in the plasma, the sensor for a mass spectrometer system is a complicated unit with its own technology. There are three basic types of mass spectrometers that have seen applications in spaceborne investigations. They include the Bennett Ion Mass Spectrometer (BIMS), the Quadrupole Ion Mass Spectrometer (QIMS), and the Magnetic Ion Mass Spectrometer (MIMS). The first two rely on mass filtering techniques that employ appropriate combinations of dc and rf fields, while the third uses magnetic deflection to separate and detect the individual ion constituents. Because of their greater flexibility, the treatment here will concentrate on the quadrupole and magnetic devices. In both cases the spectrometers can also measure the neutral mass constituents by employing an ionizing source in front of the filter unit. This can present some additional problems in ion optics, but with proper care it increases the versatility of the device.

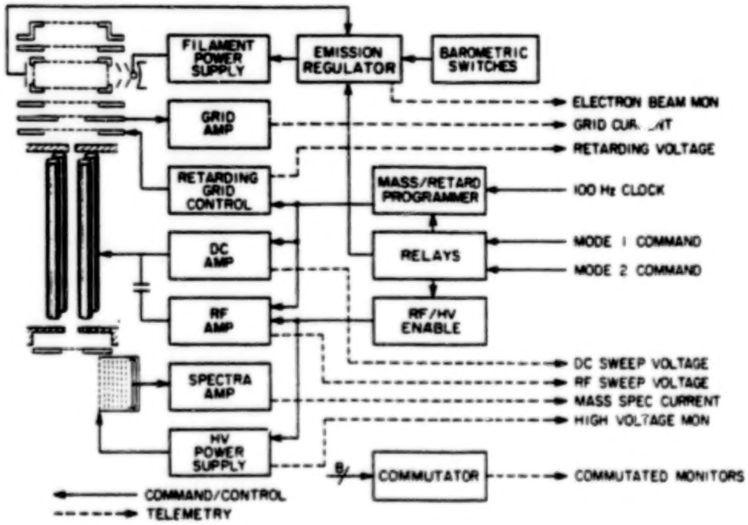
Figure 12A provides a schematic diagram of a QIMS (actually a QINMS, since an ionizer is included for neutral mass spectrometry measurements) and Figure 12B provides a block diagram of the associated electronics [see *Hunton et al.*, 1986, for instrument details]. Because of high vacuum requirements in the filtering unit, and the possibility of high-voltage arcing in the electronics unit and multiplier/sensor subassembly, the sensor package is divided into an evacuated section above the mounting flange and a sealed section below. The evacuated section of the detector is sealed (in this case) by a motor driven cover at the ion/neutral entrance-aperture end of the instrument. Other applications make use of a pyrotechnic device which expels the cover at orbital altitudes or at a proper altitude within a sub-orbital rocket trajectory.

The device shown in Figure 12A is sensitive to neutral and ionic species, though not simultaneously, and thus provides full spectrometric information on the ion and neutral constituents in the "ion mode" and "neutral mode," respectively. In addition to the heavy particle composition, the instrument provides a measurement of the total pressure of the neutral gas and the total density of the ion species in the neutral and ion modes, respectively. This information is provided by currents collected by a biased grid which is located between the ion source and the quadrupole rods. That grid collects a fraction of the total ion current leaving the ion source. Additional sensor capabilities come from a grid in the ion source that functions as a retarding potential analyzer (RPA).

In the ion mode, the ion attractor element is biased normally at -5 to -10 volts with respect to the spacecraft ground, in order to develop an attracting sheath to draw positive ions into the sensor. With the ionizer disabled, the attracted ions pass through the ion source and density analyzer sections and enter the quadrupole rod filtering element where programmed values for dc and rf fields allow either individual mass selectivity or an "integrated" ion mode (*i.e.*, all ions of mass greater than some pre-determined value can pass through the filter and be detected by the multiplier unit). The rf frequency is a constant value determined by several factors, including the desired mass range of the spectrometer. The magnitudes of the rf and dc components determine the mass of the ions that will have stable trajectories through the rods [*Paul et al.*, 1953, 1955 and 1958; *Dawson and Whetten*, 1969; *Bailey et al.*, 1982] and their amplitudes are varied in such a way that the rf/dc ratio is held constant. The device is basically a swept instrument, with either the value of the rf or dc voltages providing a monitor of the mass being sampled. A continuous sweep mode designed to cover the 1-64 amu range typically might require 0.5 seconds. Along a satellite trajectory at orbital velocities of 8 km/s, this would result in a complete spectrum having been



A. Schematic diagram of internal construction of QINMS



B. Block diagram of QINMS electronics

Figure 12. Schematic illustration of the Quadrupole Ion and Neutral Mass Spectrometer QINMS (Panel A) and an associated block diagram of Control Electronics (Panel B). [From *Hunton et al.*, AFGL-TR-86-0084 (7 April 1986)].

measured over a 4 km segment of the orbit. Faster sampling has been achieved by a digitally-programmed sweep which identifies pre-selected masses and samples each for a 10 ms period. In this format, for example, the device can step through O^+, N_2^+, NO^+, O_2^+ and Fe^+ in 50 ms, or an equivalent along-track distance of 400 meters. Developments in the detector system and its electronics suggest that mass dwell times could be reduced to values as low as 1 ms [*J. Ballanthin*, private conversation], providing an order of magnitude improvement in space/time resolution.

For purposes of comparison it is useful to introduce the MIMS, since it is intrinsically a faster device than the QIMS. This comparison is most easily accomplished with reference to the AE ion mass spectrometer [*Hoffmann et al.*, 1973] illustrated in Figure 13A. It is a magnetic deflection device relying on the Lorentz $\vec{V} \times \vec{B}$ force to separate the ions. It consists of an entrance aperture, an accelerator section, a magnetic analyzer, and a detector system. The electric and magnetic fields are so arranged as to produce a mass spectrum along a focal plane following the magnetic analyzer. Three slits are placed along the focal plane in appropriate places to simultaneously collect ions in the mass ratio 1:4:16. Behind each slit is an electron-multiplier log-electrometer-amplifier detector.

The diagram in Figure 13A shows the MIMS, with its entrance slit, magnet, drift tubes, exit slits, and electron multipliers. The entrance grid is a screen, which is positioned flush with the external spacecraft ground plane and electrically and thermally isolated from it. Directly behind the grid is a 2.5-cm diameter hole through a grounded plate and behind that is a small entrance slit, of approximately 0.01-cm^2 area, which forms the entrance aperture and ram inlet pumping port for the instrument package. Like the QIMS, the MIMS requires vacuum conditions in its mass analyzer.

The $\vec{V} \times \vec{B}$ mass discrimination is achieved in the analyzer magnet assembly, with a field strength of 3600 gauss in the gap, and a stray field reduced to less than 350γ at a distance 50 cm from the instrument-package center by the use of magnetic shielding materials.

Figure 13B is a block diagram showing the circuits which operate the analyzer, process the data, and interface with the spacecraft. The AE instrument operated without a conventional ion source, collecting only positive ions from the ionosphere. These were accelerated through the entrance aperture into the analyzer system by a negative sweep voltage usually with an exponential decay waveform from a programmed (resistance-capacitance decay) high-voltage power supply. As its voltage decreased, ions of progressively larger mass were successively focused on each of the three analyzer collector slits. In its fastest mode, the mass ranges 1:4, 4:16, and 16:64 amu were scanned simultaneously in the low-, mid-, and high-mass ranges, respectively, providing

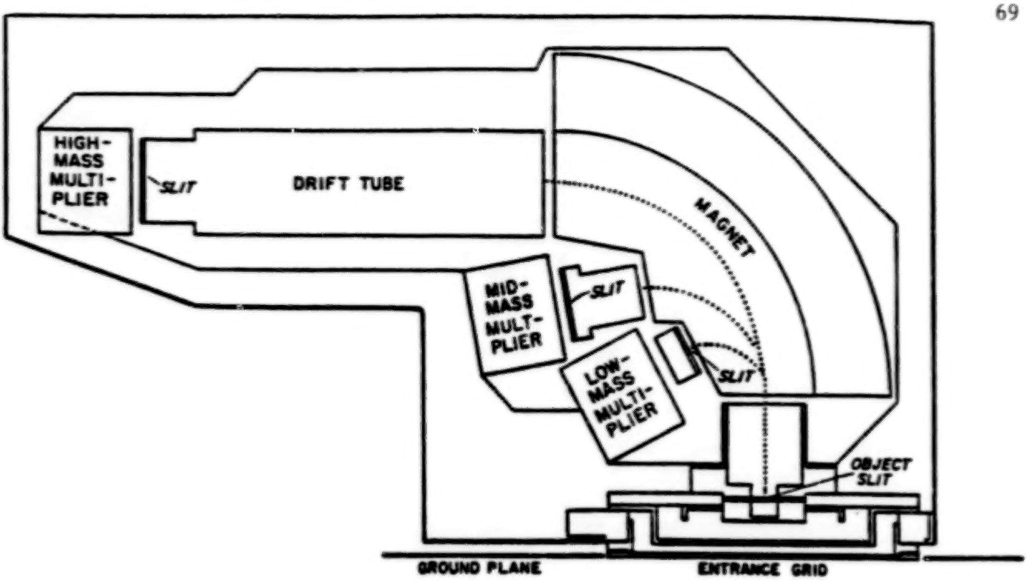


Figure 13A. Outline drawing of mass spectrometer showing entrance grid and slits, magnet with ion trajectories, collector slits, and electron multipliers. Collector slits are placed to simultaneously collect ions in the mass ratio 1:4:16. [From *Hoffman et al.*, Rad. Sci. 8, 315 (1973).]

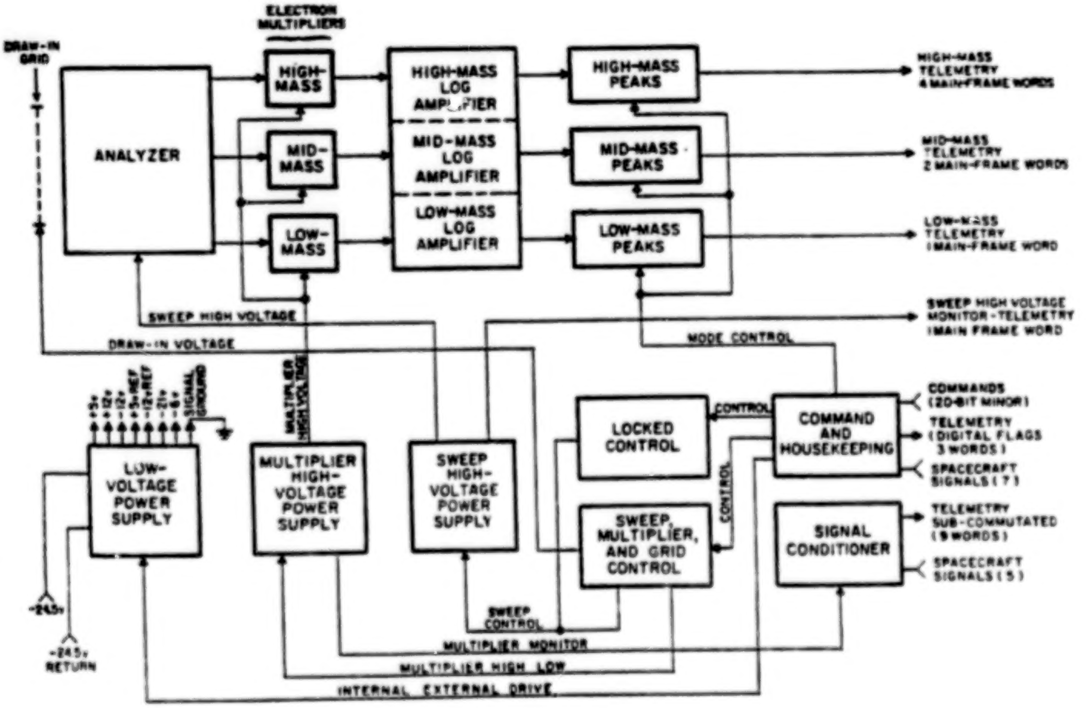


Figure 13B. Block diagram showing analyzer and electronics as well as interface with spacecraft. [From *Hoffman et al.*, Rad. Sci. 8, 315 (1973).]

the entire spectrum from 1 to 64 amu in 1 sec.

While a one sec sweep appears to be a comparatively slow rate, telemetry on the AE spacecraft was one of the major limiting factors. However, the nature of the device (not necessarily its operation on the AE spacecraft) is such that it has an intrinsically faster mass sampling capability. This faster capability is a direct consequence of the multiple-mass sensing elements arranged in a mass ratio of 1:4:16. This implies, for example, that in any 10 ms mass sampling interval, the AE-MIMS could simultaneously detect three masses. An ideal combination in the topside ionosphere would be H^+, He^+ and O^+, with the instrument providing 10 ms (80 meter) "snapshots" of the topside distribution of these light ions. This simultaneous mass sampling MIMS technique has been applied in the Giotto mass spectrometer [***Krankowsky et al.***, 1981] where five simultaneous mass ratios are sampled. The technique can be pushed even further using planar channel electron multiplier arrays for an instantaneous mass spectrum over the 1 to 64 amu range.

Retarding Potential Analyzer (RPA). The ion temperature is the remaining parameter necessary to define the complete state and condition of the ionospheric plasma. The instrument which has contributed most in specifying the thermal ion energies is the retarding potential analyzer (RPA), with a lineage going back almost 30 years [***Hanson and McKibbin***, 1961; ***Hanson et al.***, 1970 and 1973]. Applications have included rockets, several Air Force satellites, and the OGO-6, Viking, Atmosphere Explorers C, D, and E, and the DE-2 spacecraft. Initially the RPA was used to measure the ion concentration (N_i) and ion temperature (T_i). In satellite applications it can also sort N_i into its constituent parts if the various ions present have widely different masses. For example, at appropriate flux levels it could identify contributions to its collected current from O^+, NO^+ and Fe^+, but it cannot separate N_2^+, NO^+ and O_2^+, nor can it identify minor constituents. Tests with the OGO-6 data also revealed the capability for measuring the bulk ion velocity component normal to the sensor face.

Illustrated in Figure 14 is the RPA on the DE-2 spacecraft [***Hanson et al.***, 1981]. It utilizes the classical planar design which requires orientation with its front face nearly normal to the vehicle velocity vector. In this configuration, ions entering the aperture in the front face pass through a region of several gridded stages with differently applied potentials. The two entrance grids (G1) are grounded to the vehicle and surrounded by a ground plane. The next grid inward is the double retarding grid (G2) to which a time-varying voltage is applied. The suppressor grid (G3) is held at a negative potential ($\approx$ -15V) to prevent low-energy ambient electrons from reaching the collector and to prevent secondary-electron escape from the collector. The shield grid is held at spacecraft ground, and is designed to protect the collector electrometer from electrical

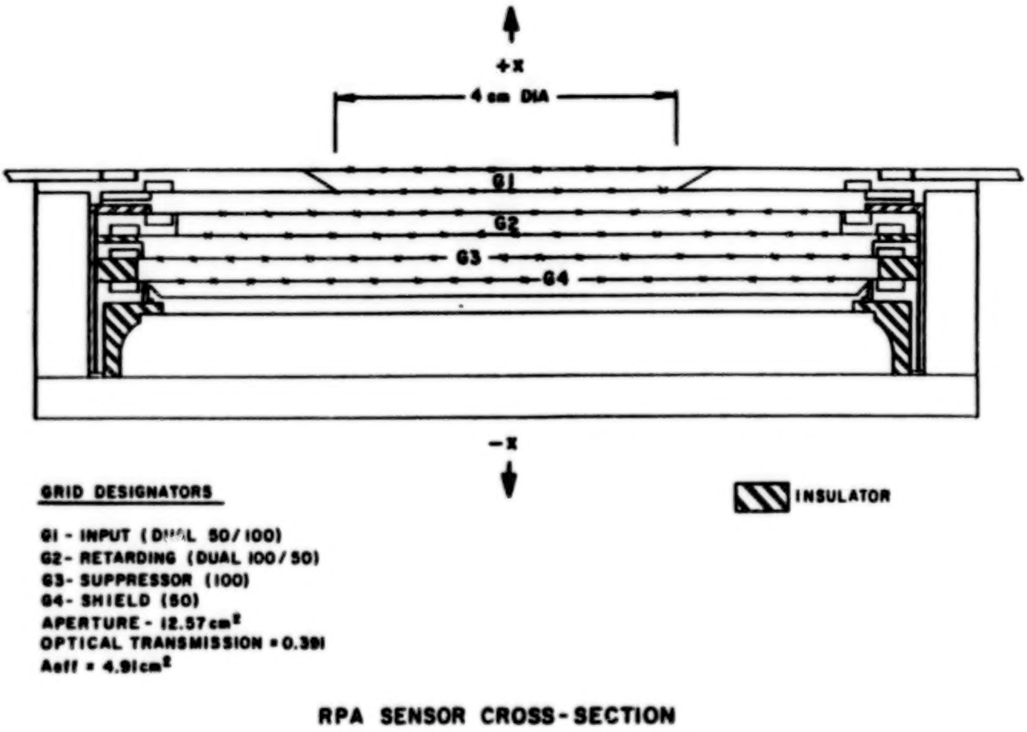


Figure 14. Schematic cross section of the DE-2 RPA sensor. The numbers in parenthesis describing the grids give the number of one mil wires per inch. The nominal grid element separation is 2.5 mm. [From *Hanson et al.*, 1981].

transients generated by changing potentials on the retarding grids.

The retarding potential (on grids G2) is variable in the range from +32 to 0 V, with sweep periods varying from 300 - 700 ms. The collected ion current corresponding to an applied retarding grid voltage is synchronously detected, amplified and filtered before being presented to the spacecraft data acquisition system. The resulting I-V characteristics represent the primary data from the RPA. Subsequently a least-squares fitting technique is used to retrieve the ion temperature, total ion concentration, information on the composition of major ions (as discussed above) and the component of the ion-drift velocity that is parallel to the sensor normal.

3.3 Techniques to Measure Thermospheric Parameters

It has been established that there is a close coupling between the neutral and ionized species in the ionosphere-thermosphere system, and accordingly measurements which define the state and condition of the charged constituents must be complemented by an equivalent set of measurements to define the thermospheric state.

The thermosphere responds dynamically to the energy and momentum sources in the solar EUV and UV wavelength regimes, to precipitating energetic particles in the auroral region, and to ion-neutral collisions. Neutrals driven by ion drag can modify the Joule heating rates which in turn can influence the neutral composition and temperature distributions and feedback into the ionosphere. Convection forces at high latitudes can drive wind systems that effect the global distribution of ionospheric electric fields. Winds in the lower ionosphere can generate dynamo fields which map equatorward along flux tubes, and these same winds can bring about complicated distributions of intermediate plasma layers at altitudes co-located with wind-shear nodes.

Prior to the AE series and DE-2 spacecraft missions, little was known about the intimate coupling between the charged and neutral species in the ionosphere-thermosphere system. In many respects the thermospheric database is still relatively primitive. We have yet to establish: (1) the seasonal, diurnal and lat/long dependencies of the thermospheric wind system, (2) the roles of the lower ionospheric dynamo under steady and dynamic conditions, (3) the roles of the diurnal and semi-diurnal tides and wind shear in controlling the plasma distributions at altitudes below 200 km, and (4) the global impact of the fossil wind system on ionospheric electric fields and associated controls of the mid- and equatorial ionosphere. ("Fossil wind" is a term coined by A. Richmond to describe thermospheric winds equatorward of a poleward retreating auroral oval.) "In situ" measurements can provide the much needed database to help resolve these issues. To understand the requirements and the complexities of the measurements, we focus here on three instruments: the Wind and Temperature Spectrometer

(WATS), the Fabry-Perot Interferometer (FPI), and the Temperature and Wind Sensor (TAWS). The FPI is an optical remote sensing technique which complements the "in situ" measurements of the WATS on the DE-2 spacecraft. The TAWS is an outgrowth of the WATS technique, designed for use on rocket and satellite investigations, and to eliminate the mass spectrometer component of the WATS configuration.

Winds and Temperature Spectrometer (WATS). The Winds and Temperature Spectrometer on the DE-2 spacecraft employs a neutral mass spectrometer as the basic diagnostic element to measure the neutral constituents and their densities relative to an upstream baffle. Currents measured in the mass spectrometer as a function of the baffle position provide information on the incident velocity vector. This adaptation of mass spectrometry to measurements of wind and temperature is a relatively recent development, first employed on the AE satellites and later in a much improved version on the DE-2 spacecraft [*Spencer et al.*, 1981 and 1982; *Wharton et al.*, 1984].

Figure 15 illustrates the basic instrument concept, showing the mass spectrometer inlet orifice and a baffle geometry. (The mass spectrometer is a quadrupole system with an antechamber to improve the accuracy of the O and O_2 density measurements. These details are not important for the description offered here, and the interested reader is referred to detailed instrument description in *Spencer et al.* [1981].) The baffle is moved across the instrument input at a distance several centimeters upstream of the input orifice. The baffle movement is done at a precisely determined speed during measurements of the abundance of the major thermospheric constituents, such as O and N_2. The exact angle at which the incoming beam of neutral particles experiences greatest reduction at the input orifice to the mass spectrometer is a function of the cross-track wind component in the plane of the baffle motion. The DE-2 WATS device employed two orthogonal baffle configurations allowing determination of the vertical and zonal (*i.e.* components horizontal and perpendicular to the orbit plane) thermospheric wind components. Shown at the right in the figure is an illustration of an observed N_2 density profile in the mass spectrometer during one pass of the baffle in front of the orifice. This instrument response comprises the basic data of the wind measurement technique. The density of only a single species is considered in the measurement procedure. That procedure focuses on the amplitude variation of density with time due to baffle modulation of the entering stream as illustrated in Figure 15. The direction of arrival of the particles, and thus the signal minimum, occurs at the point in the scan where the stream, baffle, and orifice are colinear. Taking into account the velocity and orientation of the spacecraft and the measured angle between the orifice normal and direction of the minimum, permits calculation of the wind component lying in the plane defined by satellite velocity V_{sat} and the baffle scan path. The shape of the "bite out" portion of the density curve is determined by the velocity of the spacecraft, the

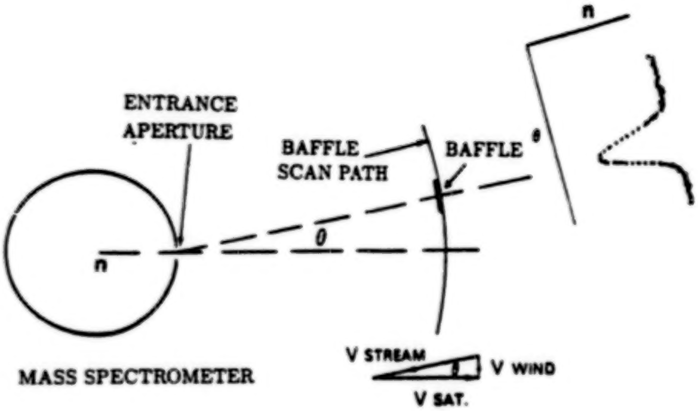


Figure 15. Schematic for the DE WATS measurement of thermospheric wind and temperature. The aperture of the closed-source mass spectrometer is scanned systematically in the vertical and horizontal planes by two baffles. (Only one baffle scan is illustrated.) As the baffle scans in front of the aperture, the measured signals drop with a characteristic signature as shown by the $n(\theta)$ plot. The baffle location at the time of maximum signal drop-off is related to the wind component in the plane of the baffle motion, and this wind can be obtained through solution of the vector triangle shown. [Adapted from *Killeen and Roble*, 1988 and *Spencer*, private communication, 1989].

scanning speed, the geometry of the baffle-orifice system, and the transverse velocity distribution (and therefore the temperature) of the N_2 particles. The gas temperature is therefore calculated from this observed density variation.

As discussed, the thermospheric wind can be obtained by solving the vector triangle shown in Figure 15. That solution requires accurate knowledge of the spacecraft velocity vector, which is obtained from a combination of horizon, Sun, and star sensor data. In the final analysis, the spacecraft velocity vector is considered to be known with sufficient precision to provide WATS wind measurements with an accuracy approaching ±15 m/s [*Killeen and Roble*, 1988].

Temperature and Wind Sensor (TAWS). While the TAWS is an outgrowth of the AE and DE-2 instruments of *Spencer et al.* [1981], it represents a completely different design which uses no moving parts, does not require a mass spectrometer, and measures the energy distribution using the established technique of electrostatic deflection [*Hanson and Heelis*, 1975; *Heelis et al.*, 1983]. The TAWS has been developed at NASA/GSFC (F. Herrero, private communication) and has had recent successes on two rocket-borne investigations. Because of its "unpublished" status, we offer a more graphic explanation of its operating principles.

Like WATS, the temperature and wind vector of the neutral atmosphere are obtained from TAWS measurements of the angle and energy distributions of the drifting Maxwellian atmospheric stream in the moving frame of the spacecraft (see Figure 16A for an overview of the measurement concept). A small aperture on the ram surface of the instrument accepts the incident stream of gas which moves supersonically with velocity $V_R = U + V_S$, V_S being the spacecraft velocity and U the wind vector. In Figure 16A ϵ defines the angle between the surface normal n and the velocity vector V_S, and θ defines the angle between V_R and V_S. Using an array of detectors which measures the angular distribution of incident particles along an arc gives $U_\perp$, the wind component perpendicular to V_R lying in the plane of the detector arc. The component $U_\parallel$ may be obtained from the energy distribution of the incident stream, in a manner similar to retarding potential analysis [*Hanson and Heelis*, 1975]. A second detector array with its arc perpendicular to the first is used to obtain the third component of the wind vector. The measured energy and angle distributions are fitted to theoretical distributions in order to obtain the wind velocity and temperature. The satellite attitude enters into the determination of $U_\perp$ through the angle ϵ in Figure 16A, with post-flight analysis of the data required to obtain an accuracy of 15-30 m/s. This represents a systematic error in $U_\perp$ only, the resolution being independent of the altitude error. $U_\parallel$, N_n and T_n are obtained directly from the distributions and are insensitive to an error in attitude as long as the instrument points within a few degrees of the ram direction (F. Herrero,

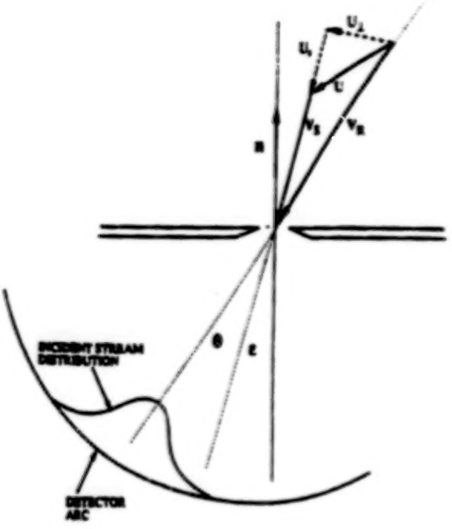


Figure 16A. Schematic description of the incident stream and the angular distribution produced in the detector arc when the satellite velocity V_S adds with the neutral wind U to give a net vector V_R at an angle θ with respect to V_S. [From *F. Herrero*, private communication.]

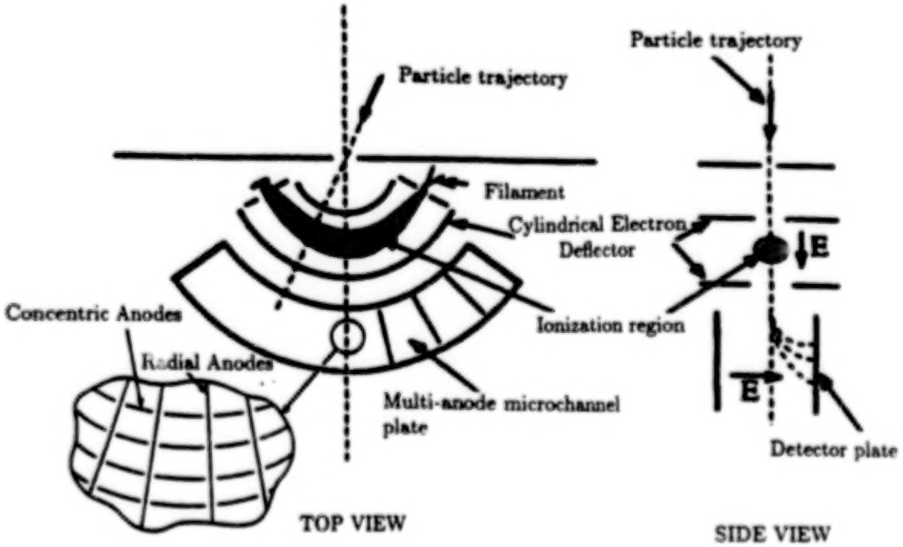


Figure 16B. Illustration of the TAWS Detector. [From *F. Herrero*, private communication.]

private communication).

Figure 16B provides a more detailed illustration of the TAWS. A 0.040 inch entrance aperture is preceded by a pair of shielded plates (not shown) which keep the ambient ions and electrons from entering the spectrometer. Once inside the detector, the neutral stream intersects an electron beam which is moving along a cylindrical electrostatic field. In that region a fraction of the neutral atoms and molecules are ionized, with some of the molecules undergoing dissociative ionization as well. In addition to maintaining the electron beam along the desired arc, the cylindrical electrostatic field withdraws the ions produced there, preserving the number density of ions at the corresponding angle of incidence. This is particularly important in contending with ions produced via dissociative ionization, which may be initially projected in a direction perpendicular to the incident molecular ray with an energy of several eV. For these ions, the radial electrostatic field rapidly deflects them in a direction closely parallel to the initial molecular ray at the corresponding angle θ. As the side view of the figure shows, the ions withdrawn from the ionization region are deflected by a constant electric field onto the detector plate. The detector is a microchannel plate with a CODACON anode structure beneath to provide the angle and energy analysis. The multiple radial anode structure shown in the inset allows 1° angular resolution over an angular range of 90° as shown. The energy analysis is achieved by the concentric arc anodes shown beneath the radial anodes in the inset. The energy resolution $\Delta E/E$ is a function of the radial position of the anode arc and is optimum for the median arc, never exceeding 0.15 based on the initial energy before acceleration out of the cylindrical field. This is sufficient in view of the natural energy spread of the incident stream. Because of the drawout field, the energy range of the analyzer extends from 0 to 20 eV. The contact potential in the analyzer may be obtained as part of least squares fit to the multiple mass peaks obtained in routine energy spectra. A TAWS flight configuration would have two mutually orthogonal sensor heads, for a full vector determination of the neutral wind velocity along the satellite track or along the trajectory of a sub-orbital rocketborne investigation.

The Fabry-Perot Interferometer (FPI). The "in situ" wind and temperature measurements of the WATS and TAWS instruments can be greatly augmented by companion remote sensing capabilities of a satellite-borne Fabry-Perot Interferometer (FPI). This was the case in DE-2 where a FPI measured altitude profiles of winds and temperatures ahead of and below the spacecraft by remotely sensing the Doppler shift and Doppler line broadening of the $O(^1D)$ thermospheric emission line profile at 630 nm [***Killeen and Roble***, 1988]. The instrument field of view on DE-2 is directed along and below the spacecraft velocity vector (Figure 17), and horizon scans of the Earth's limb are performed in the orbit plane with an internal scanning mirror. The measured

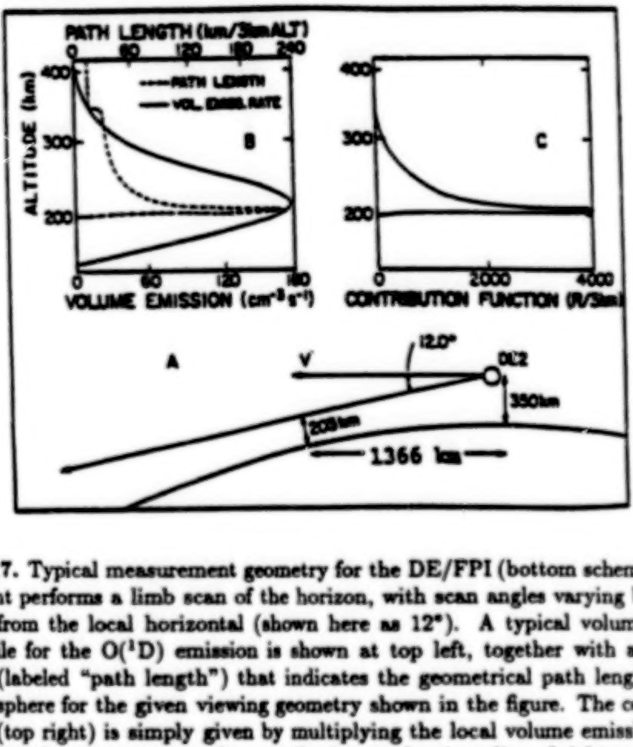


Figure 17. Typical measurement geometry for the DE/FPI (bottom schematic). The instrument performs a limb scan of the horizon, with scan angles varying between 5° and 15° from the local horizontal (shown here as 12°). A typical volume emission rate profile for the O(^{1}D) emission is shown at top left, together with a weighting function (labeled "path length") that indicates the geometrical path length through the atmosphere for the given viewing geometry shown in the figure. The contribution function (top right) is simply given by multiplying the local volume emission rate by the path length and represents the contribution to the given line-of-sight measurement from the various altitude layers. [From *Killeen and Roble*, 1988.]

Doppler shift is converted into a line-of-sight neutral wind measurement by subtracting the component of the shift due to the spacecraft velocity; and the kinetic temperature is detemined from the measured signal after subtracting the instrumental broadening function. This allows the temperature to be determined from characterization of the Gaussian half width of the emission line [***Hays et al.***, 1981; ***Killeen and Hays***, 1982; ***Killeen and Roble***, 1988].

The FPI measurements are spatially integrated in the (meridional) viewing direction, typically over ~ 500 km or so, and the results are usually referred to the location of the tangent point along the line of sight of the instrument (see bottom panel, Figure 17) with the "contribution function" (shown in the top right panel of Figure 17 for a typical case) describing the altitudinal origin of the optical emission detected by the FPI for a given viewing geometry. For tangent point altitudes greater than the altitude of the $O(^1D)$ emission maximum, the dominant contribution to a given measurement comes from altitudes near the tangent point altitude.

While the spatial and temporal resolution of the FPI is dictated by scan time and by the integrated-field-of-view nature of the data, it represents a significant augmentation to the wind measurement of WATS. This augmentation was manifested by an analysis technique developed by ***Killeen et al.*** [1982] that merges the data from both instruments to provide a composite description of the horizontal neutral wind vector along the orbital track of DE-2. The technique extrapolates the remotely-sensed altitude profile of the FPI meridional wind measurements upward to the altitude of the spacecraft; and the extrapolated FPI results are merged with the "in situ" WATS zonal wind measurements to provide the summed horizontal vector wind. This superposition of data proved invaluable in mapping important phenomena in the ionosphere-thermosphere system, including geomagnetic control of winds in the polar region, characterization of the circulation patterns in the central polar cap, and ion-neutral coupling [***Killeen and Roble***, 1988; ***Killeen et al.***, 1984a and b; ***Killeen et al.***, 1988].

3.4 Electric and Magnetic Fields, Energetic Particles and Plasma Boundaries

In preceding sections, we have developed a baseline instrument complement capable of defining the state and condition of the ionospheric-thermospheric system. To understand those observations and fold them into the bigger picture of cause-effect relationships, we need to consider incident radiation in the UV and EUV portions of the solar spectrum, electric and magnetic fields, energetic particle inputs and plasma boundaries. We concentrate on the "in situ" measurements of energetic particles and controlling fields.

Electric Fields. Electric fields are central to the issue of "electrodynamic coupling," with the roles of the thermosphere and magnetosphere at the heart of the problem. A convenient (but admittedly oversimplified) view of ionospheric electric fields would identify two morphological/phenomenological domains: (1) the high-latitude domain poleward of the shielding layer (generally viewed as being near the equatorward boundary of the diffuse aurora), and (2) the remainder of the ionosphere equatorward of the auroral oval (which we refer to as the mid- and equatorial latitude region). Under quiet conditions, the magnetosphere dominates the electric field patterns at high latitudes, with ion-neutral coupling not insignificant, while the thermospheric dynamo dominates the mid- and equatorial regions, with little magnetospheric contribution. Under dynamic conditions however, the roles are intertwined at all latitudes, with no current database capable of sorting out the contributions.

The existing database on electric fields comes primarily from "in situ" differential floating potential measurements on extended booms, and from "in situ" and radar ion-drift measurements tied to $\vec{E} \times \vec{B}$ velocity effects [see *e.g. Heelis et al.*, 1981; *Heelis*, 1988]. In some regards, high-latitude electric fields have been better diagnosed and are more fully understood, but controversy still exists. Mid- and equatorial electric fields are not as well understood, with present specifications relying primarily on an empirical model developed by the incoherent radar chain in the 75° meridian [*Richmond et al.*, 1980]. Along with ground-based radars, important contributions to the current database for global electric field patterns (and the roles of the thermosphere and magnetosphere) have come from the DE investigation [see *e.g. Heelis*, 1988]. At high latitudes DE clearly showed that the time history of the field-driven convecting ions and the dynamics of the neutral gas were important to understanding the chemistry, dynamics and energy balance of the system. Unfortunately, the local time and season of the observations were locked together in a 90° inclination orbit, leaving a restricted data set and a scientific need for more comprehensive global-scale measurements.

It is clear that knowledge of the ambient electric field is vital to understanding and interpreting plasma motions, waves and turbulence at both large and small scales. DC electric field measurements are useful for studying large scale convection patterns, while wave electric fields are important in plasma instability and turbulence studies. Accordingly, electric field probes have had a long and successful flight history, beginning with simple DC instruments flown on sounding rocket and balloon payloads in the sixties [*e.g. Johnson and Kavadas*, 1963; *Mozer and Serlin*, 1969], and extending into the seventies and eighties with global-scale satellite investigations [*Maynard et al.*, 1988, *Hoffman et al.*, 1988 and *Heppner and Maynard*, 1987] and complex broadband wave analyzers in the present era of rocket-borne investigations [*e.g. Kellogg et al.*, 1986; *Kelley and Earle*, 1988].

The floating double-probe technique is by far the most common method for "in situ" electric field measurements on satellite and rocket-borne payloads, so it will be the focus of this discussion. Since only refinements in the electronics system have been made in the method since the early flight tests, much of the basic developmental work [***Fahleson***, 1967; ***Mozer***, 1973] is still quite relevant today. The double-probe technique uses high-impedance difference amplifier circuits to measure the difference in the local floating potential between two small conductive probes (typically spherical in shape) that are extended on long booms from the rocket or satellite body. Since the floating potential and the local space potential in a plasma are related by a function involving the ion mass, electron temperature, and the ratio of the probe radius to the Debye length [***Szuszczewicz***, 1972] the measured floating potential difference between identical probes is proportional to the spatial electric field between them. This assumption is valid only when the ion mass, electron energy and Debye lengths are the same at each probe. These conditions are generally guaranteed to prevail in quiescent, laminar ionospheric domains, and in disturbed regions where irregularity scale sizes are much greater than the double-probe separation.

The assumption of similar plasma distributions near the two probes is implicit in the interpretation of all floating double probe measurements, and it highlights a potential limitation in dynamic, energetic and highly irregular plasma environments. Despite this necessary assumption, however, the floating double probe technique has proved extremely useful in diagnosing high- and low-latitude ionospheric electric fields [***Heppner and Maynard***, 1987; and ***Maynard et al.***, 1988] and geophysical processes as diverse as two-stream and gradient-drift waves in the electrojet regions [***Pfaff et al.***, 1982; 1984], convection velocity shears [***Kelley and Carlson***, 1977; ***Basu et al.***, 1988; ***Earle et al.***, 1989], spread-F irregularities [***LaBelle et al.***, 1986], and lightning induced transients in the ionosphere [***Holzworth et al.***, 1985].

While measuring the potential difference between two probes is conceptually very simple, there are a variety of complications and subtleties involved in properly performing and interpreting the measurements. For example, the input impedance of the measuring circuit should be as large as possible with respect to the resistance of the plasma sheath which surrounds the probe, so that the probe electronics do not provide alternative current paths to the plasma. Input impedances of 10^{12} Ohms or more are common in modern instrumentation. It is also essential that the pair of probes used in the difference measurement be as similar as possible in size, shape, composition and distance from the spacecraft, since dissimilar sheaths, probe surface work functions, levels of surface contamination, contact potentials, wakes or current responses can introduce large errors [***Fahleson***, 1967]. In addition to using identical probes, each probe should be symmetric with respect to the boom on which it is mounted, so that shadowing does not asymmetrically affect the photoelectron current to the probe at different portions of its spin cycle. For this reason a small tubular section of the boom is typically extended out beyond each probe, in order to cast an appropriate shadow and create symmetry with the opposing probe [*Pedersen et al.*, 1978].

By using multiple probe sets, as shown in Figure 18, the electric field in three orthogonal directions can be measured simultaneously, thereby yielding the 3-dimensional vector electric field. In practice, however, it is often difficult to measure the electric field parallel to the ambient magnetic field, since the parallel field is generally quite small in the ionosphere and the vehicle surface can interfere with the measurement when it intersects a field line which also intersects the probe [*Fahleson et al.*, 1970]. Perpendicular electric fields in the ionosphere are typically much larger, and therefore are more easily measured.

Despite these and other complexities in analysis and probe design, the floating double-probe system offers several advantages over other types of electric field instruments, such as field-mills and drift meters. One such advantage of double-probes involves the finite boom length of spinning systems which can help in determining the wavelength and propagation direction of narrowband plasma wave modes, through an interference effect known as "fingerprints" [*Temerin*, 1979; *Pfaff*, 1986]. Still another advantage exploits the Doppler shift created by the relative motion between plasma waves and the spacecraft. The Doppler shift in frequency provides phase velocity information in cases where a wave mode has a well defined fundamental frequency.

Magnetic fields. The measurement of magnetic fields is introduced into this description of "in situ" techniques not in the sense of it being a scientific objective in its own right but more in the sense of providing measurements that support other "in situ" techniques and in the sense that it aids in the overall analysis of results. Appropriate to this role is a three-axis fluxgate magnetometer with a dynamic range of ± 50,000 nT and an accuracy of 0.75 nT. At a slow sampling rate (*e.g.*, 1-5 samples/sec on each axis) the magnetometer would provide vector magnetic field measurements along the satellite track. This measurement supplies the required B-field reference for the electric field results and a corresponding baseline for $\vec{E} \times \vec{B}$ and $\vec{V} \times \vec{B}$ analysis of plasma and thermospheric wind motion and dynamo fields. At a higher sampling rate (*e.g.*, at 100 samples/sec on each axis) the instrument would provide indications of current system boundaries, a diagnostic of geomagnetic disturbances, and a monitor of Alfven and cyclotron waves (O^+ cyclotron frequency is 32 Hz in a 0.35 Gauss field). The best configuration for the magnetometer is one that allows it to be boom mounted to minimize the influence of stray fields from spacecraft systems.

Energetic particles. Next to UV and EUV portions of the solar electromagnetic

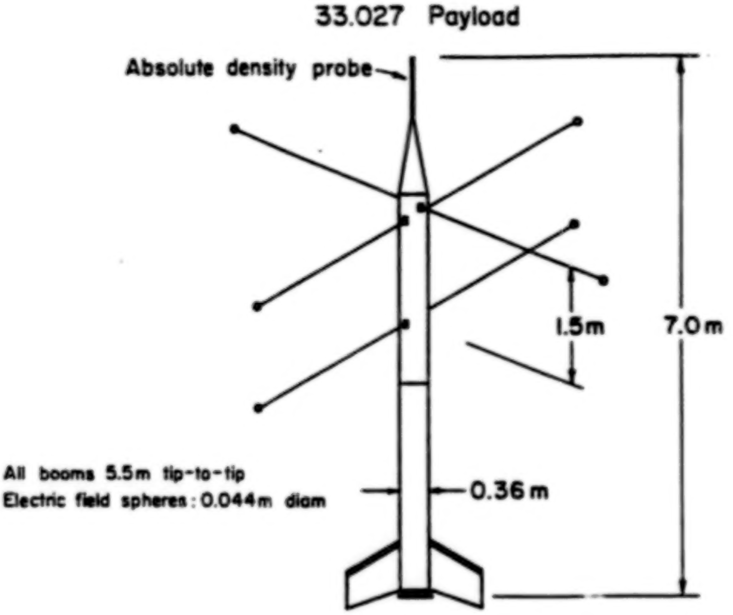

33.027 Payload
Absolute density probe
1.5m
7.0m
All booms 5.5m tip-to-tip
Electric field spheres: 0.044m diam
0.36m

Figure 18. Sketch of payload sensor configuration drawn to scale, for including second stage motor. [Adapted from *Pfaff*, 1986.]

radiation, one of the most significant energetic inputs to the ionospheric-thermospheric system involves precipitating electrons at high latitudes. Measurement of the energy spectrum and the angular distribution of velocities with respect to the magnetic field provides the three-dimensional velocity space-distribution function for deriving the energy and momentum coupling into the ionospheric-thermospheric system.

In the past, the angular distributions of precipitating particle densities and velocities were measured by a single, narrow field-of-view analyzer on a spinning platform or by arrays of fixed, narrow field-of-view analyzers, each making a different angle with respect to the local magnetic field. Although these approaches met with reasonable success, the angular, spatial and temporal information was incomplete, and the design, packaging, and calibration of the analyzers was a difficult and costly process. For example, the plasma instrument on DE-2 required 15 collimation systems, 15 sets of deflection plates, and 15 sensors to provide the energy-angle information [*Sablik et al.*, 1988]. We describe here a "top hat"electrostatic analyzer [*Sharber et al.*, 1988], also referred to as an Angle Resolving Energy Analyzer (AREA), which alleviates this problem by using a single set of deflection plates to accept particles over a 360° field-of-view in a single plane. An early attempt at such an instrument was the quadrispheric-type analyzer, flown on the Pioneer 6 and 7 interplanetary probes [*Smith and Day*, 1971] and on the ISEE 1 and 2 satellites [*Frank et al.*, 1978]. Improvements over the quadrisphere design were achieved in the analyzers developed for applications on the AMPTE mission [*Johnstone et al.*, 1985] and on the Giotto and Suisei missions to Halley's Comet [*Johnston, et al.*, 1986; *Mukai and Miyake*, 1986].

When viewed as a symmetric quadrisphere, the top-hat instrument has its roots in a Soviet device [*Mel'nikov et al.*, 1965] flown on Cosmos 12 in 1962. Initial flight development of the 360° top-hat analyzer was accomplished by a collaborative team of investigators [*Carlson et al.*, 1983; *Paschmann et al.*, 1985] from the United States and Europe. The top-hat analyzer has now been successfully flown on the AMPTE spacecraft [*Paschmann et al.*, 1985], the Giotto mission to Halley's Comet [*Reme et al.*, 1986], and a number of sounding rockets [*e.g.*, *Moore et al.*, 1986].

The "top-hat" electrostatic analyzer is a variation on the now commonplace spherical analyzer. The feature from which it gets its name is the small spherical section above the hemispherical deflection plates shown in the cross-sectional view of Figure 19. A charged particle entering between this top-hat and the upper hemispherical plate will be deflected into the region between the hemispherical deflection plates where (with appropriate charged-particle energies and applied fields) it will undergo "perfect" constant-radius deflection and strike the sensor (an array of eleven 16° azimuth sectors equally spaced around 360° in an equatorial acceptance plane). The device is axially

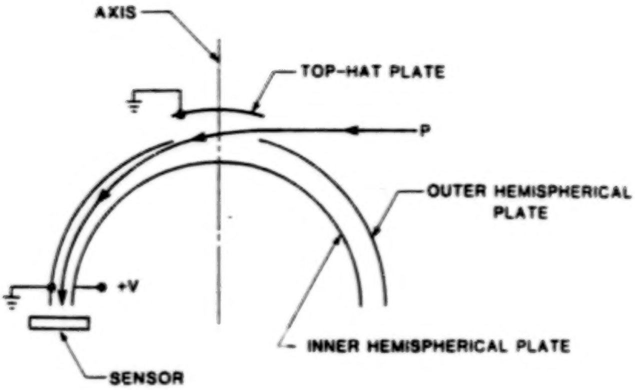


Figure 19A. Cross-sectional view of the "top-hat" electrostatic analyzer (without collimator) showing the plate arrangement and a typical particle trajectory P. For electron detection the inner hemispherical plate has deflection potential $+V$; the outer hemispherical plate and top-hat plate are grounded. [From *Sharber et al.*, 1988.]

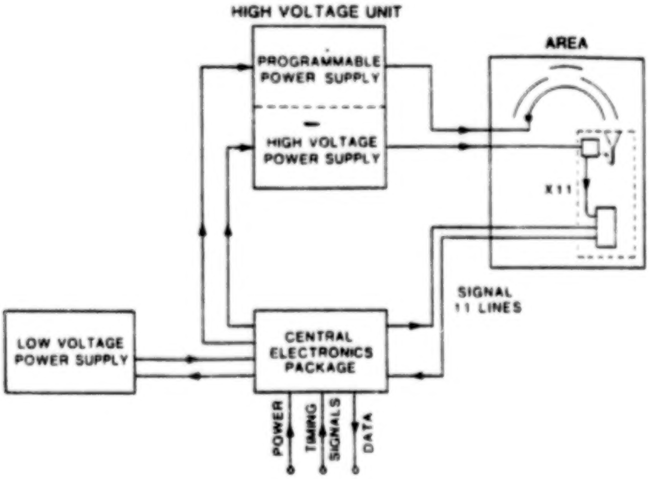


Figure 19B. Functional schematic diagram of the complete experiment showing relationships of the AREA "top-hat" deflecting and sensing unit, high-voltage unit, the low-voltage power supply, and the central electronics package. [From *Sharber et al.*, 1988.]

symmetric and therefore can provide simultaneous measurement of particles from a full 360° in the plane perpendicular to the instrument axis.

A layout of the top-hat/AREA deflection and sensor system is shown in Figure 20. The top view shows the positioning of the 11 channel electron-multiplier sensors, while the side view shows the collimator, deflection plates, and location of a channel multiplier, its amplifier, and its high-voltage biasing network. The 360° entrance aperture is defined by a series of knife-edged baffles located on the outer hemispherical plate and the top-hat plate. The baffles collimate the entrance aperture to $\pm 7.5°$ with respect to the normal to the analyzer axis.

With regard to operation and control we make reference to Figure 19B. The high-voltage unit, illustrated in that figure, contains the high-voltage and programmable power supplies. The high-voltage supply provides the Channeltron bias voltage and the programmable power supply steps the voltage on the inner hemispherical plate from 2187 to 0.200 V in 59 logarithmic steps (resolving electron energies from ≈ 10 eV $\rightarrow \approx 13$ keV). A time of 20 ms is required to reach and stabilize on the highest voltage and a time of 236 ms (4 ms/step) is required to cover the 59 logarithmic steps. On each step the settling time is 1.4 ms and the data accumulation time is 2.6 ms. A complete energy spectral measurement is made every 256 ms, or about every 2 km along the track of a low-Earth-orbiting satellite.

In this type of instrument, the energy and angular measurements suffer some degradation at the low energy portion of the spectrum (generally at energies < 50 eV). Several factors contribute to this, including contact potentials which introduce errors (uncertainties) in the deflection plate voltages for low particle energies, sheath effects in and around the entrance aperture which can modify the low-energy-particle orbits, and magnetic field effects which introduce Larmour orbits comparable to the radius of curvature in the deflection plate system. This latter effect has been mitigated in the top-hat AREA design by the use of mu-metal shielding [*Sharber et al.*, 1988].

If included on a polar orbiting satellite, an energetic particle detector like the "top-hat" design would provide baseline energy and momentum data for understanding high-latitude inputs to the ionosphere-thermosphere system. Covering energies from ≈ 10eV to ≈ 13 keV, it would provide full spectral information on major particle sources for energy deposition down to altitudes near 95 km, specification of the soft particle spectrum (in the polar cap), and definitive measurements of auroral oval boundaries.

4. Comments and Conclusions

While there is little doubt that an "in situ" database can provide a nearly-complete specification of cause-effect terms that define the state and condition of the ionosphere

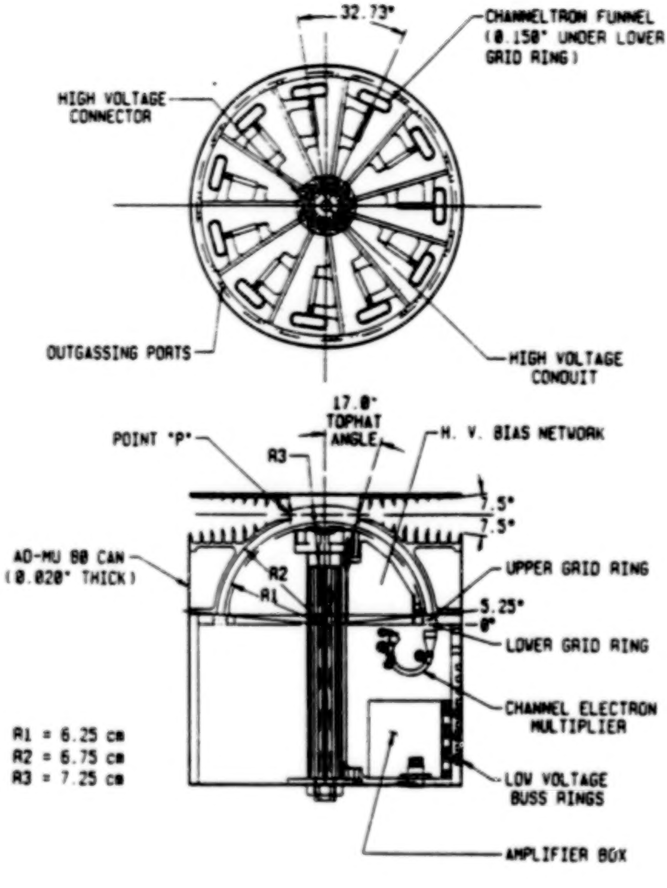


Figure 20. Layout of the "top-hat" Angle Resolving Energy Analyzer (AREA) deflection and sensing unit. [From *Sharber et al.*, 1988.]

and thermosphere, a single satellite or sub-orbital payload is limited in its ability to uniquely separate space and time variables, and develop a three-dimensional diagnosis of the interacting geospace domains. A typical approach to mitigate this problem utilizes large synoptic databases as are normally available in long-lived satellite missions. Such an approach can identify trends and ultimately specify quiescent and laminar states of global scale morphologies. More difficult, however, is the analysis of dynamic events, but there too the possibility exists for accumulating enough similar events to develop an intelligent separation of space/time controls, and integrate the results to form a three-dimensional picture. Another frequently discussed approach for future missions suggests the utilization of satellite clusters; and yet a third plan involves coordinated measurements with ground-based and "in situ" techniques. The latter possibility can be very powerful, and perhaps holds the key to solving the problem of a grossly inadequate database to define the lower ionospheric domain, its dynamo effects on the global ionosphere and its critical role in the Earth's electrodynamic circuit. The issue of intermediate layers is a case in point. Current ground-based techniques are not capable of unfolding all cause-effect terms and there are so few radar sites that a global picture cannot be developed on UT/LT, lat/long and geomagnetic controls. Satellites cannot provide a synoptic perspective in this region, and repeated multiple-site rocket launches are a financial impossibility. The solution, at least in the near term, is a coordinated ground-based and "in situ" measurement program to develop and test global-scale models that attempt to specify and predict the full system and its subsystem interactions. This is not a novel approach but one that needs to be more fully exploited to significantly advance the understanding of the ionosphere-thermosphere system and specify it as a single chemically-active and electrodynamically-responsive fluid.

Acknowledgement

I wish to extend sincere thanks to J. D. Winningham, J. Hoffman, L. Brace, W. Hanson, R. Heelis, T. Killeen, F. Herrero, J. Ballanthin and G. Earle for their inputs to the instrument descriptions. In many cases, their inputs and modified descriptions adopted from their published works were incorporated directly into the text.

References

Akasofu, S.I., Auroral arcs and auroral potential structure, in ***Physics of Auroral Arc Formation***, eds. S. I. Akasofu and J.R. Kan, AGU, Washington, D.C., ***Geophys. Monogr. Ser.***, ***25***, 1, 1981.

Akasofu S.I. and S. Chapman, ***Solar Terrestrial Physics***, Oxford University Press London, 1972.

Bailey, A., R. Narcisi, L. Wlodyka, G. Federico, and E. McKenna, Development of an ion mass spectrometer and sounding rocket system for D-region cluster-ion measurements, ***AFGL-TR-82-0022***, AD A118752, 1982.

Banks, P., W. Roberts, and J. Kropp (eds.), Solar-Terrestrial Science Strategy Workshop, Proc. NASA/MSFC Workshop, Stanford University (September 12-16, 1988).

Baron, J.J., C.J. Heinselman and J. Petrioeks, Solar cycle and seasonal variations of the ionosphere observed with the Chatanika incoherent scatter radar, ***Rad. Sci.***, ***18***, 895, 1983.

Basu, Su., Sa. Basu, E. MacKenzie, P.F. Fougere, W.R. Coley, N.C. Maynard, J.D. Winningham, M. Sugiura, W.B. Hanson and W.R. Hoegy, Simultaneous density and electric field fluctuation spectra associated with velocity shears in the auroral oval, ***J. Geophys. Res.***, ***93***, 115, 1988.

Bering, E.A., Apparent electrostatic ion cyclotron waves in the diffuse aurora, ***Geophys. Res. Lett.***, 10, 647, 1983.

Brace, L.H., B.M. Reddy and H.G. Mayr, Global behavior of the ionosphere at 1000 km altitude, ***J. Geophys. Res.***, ***72***, 265, 1967.

Brace, L.H. and R.F. Theis, Global empirical models of ionospheric electron temperature in the upper F-region and plasmasphere based on "in situ" measurements from the atmospheric Explorer-C ISIS-1 and ISIS-2 satellites, ***J. Atm. Terr. Phys.***, ***43***, 1317, 1981.

Brace, L.H., The global structure of ionospheric temperature, ***12th Plenary Meeting of COSPAR (NASA X-621-69-158)***, Prague, 1969.

Carlson, C.W., D. W. Curtis, G. Paschman and W. Michael, An instrument for rapidly measuring plasma distribution functions with high resolution, ***Adv. Space Res.***, ***2***, 67, 1983.

Chen, F.F., *Plasma Diagnostic Techniques*, ed. by R.H. Huddlestone and S.L. Leonard, Academic, New York, 1965.

Curtis, S.A., W.R. Hoegy, L.H. Brace, N.C. Maynard, M. Sugiura, and J.C. Winningham, DE-2 cusp observations: role of plasma instabilities in topside ionospheric heating and density fluctuation, *Geophys. Res. Lett.* *9*, 997, 1982.

Curtis, S.A., W.R. Hoegy, L.H. Brace and J.C. Winningham, Cusp altitudinal electron temperature gradient: DE-2 implications for heating mechanisms, *J. Geophys. Res.*, *90*, 4415, 1985.

Dawson, P.H. and N.R. Whetten, Mass spectroscopy using rf quadrupole fields, *Adv. in Electronics and Electron Phys.*, *27*, 59, 1969.

Earle, G.D., Electron-temperature determination in space plasmas with sounding-rocket electric field booms, *Rev. Sci. Instrum.*, *60*, 3076, 1989.

Earle, G.D., M.C. Kelley and G. Ganguli, Large velocity shears and associated electrostatic waves and turbulence in the auroral F-region, *J. Geophys. Res.*, *94*, 15321, 1989.

Fahleson, U., Theory of electric field measurements conducted in the magnetosphere with electric probes, *Space Sci. Rev.*, *7*, 238, 1967.

Fahleson, U.V., M.C. Kelley and F.S. Mozer, Investigation of the operation of a d.c. electric field detector, *Planet. Space Sci.*, *18*, 1551, 1970.

Fahleson, U.V., C.-G. Fälthammar and A. Pedersen, Ionospheric temperature and density measurements by means of spherical double probes, *Planet. Space Sci.*, *22*, 41, 1974.

Fontheim, E.G., L.H. Brace and J.D. Winningham, Properties of low energy electron precipitations in the cleft during periods of unusually high ambient electron temperature, *J. Geophys. Res.*, *92*, 267, 1987.

Frank, L.A., D.M. Yeager, H.D. Owens, K.L. Ackerson and M.R. English, *IEEE Trans. Geosci. Electron.* *GE-16*, 221, 1978.

Grebowsky, J.M. and N. Reese, Another look at equatorial metallic ions in the F-region, *J. Geophys. Res.*, *94*, 5427, 1989.

Gurnett, D.A., S.R. Mosier and R.R. Anderson, Color spectrograms of very-low-frequency Poynting flux data, *J. Geophys. Res.*, *76*, 3022, 1971.

Hanson, W.B. and R.A. Heelis, Techniques for measuring bulk gas-motions from satellites, *Sp. Sci. Instrum.*, *1*, 493, 1975.

Hanson, W.B., R.A. Heelis, R.A. Power, C.R. Lippincott, D.R. Zuccaro, B.J. Holt, L.H. Harmon and S. Sanatani, The retarding potential analyzer for dynamics Explorer B, *Sp. Sci. Instrum.*, *5*, 1503, 1981.

Hanson, W.B. and McKibbin, *J. Geophys. Res.*, *66*, 1667, 1961.

Hanson, W.B., S. Sanatani, D.R. Zuccaro and T. Flowerday, *J. Geophys. Res.*, *75*, 5483, 1970.

Hanson W. B., D. L. Sterling and R. F. Woodman, Source and identification of heavy ions in the equatorial F-layer, *J. Geophys. Res.*, *77*, 5530, 1972.

Hanson, W.B., D.R. Zuccaro, C.R. Lippincott and S. Sanatani, The retarding potential analyzer on atmospheric explorer, *Rad. Sci.*, *8*, 5483, 1973.

Harper, R. M., Tidal winds in the 100-200 km region at Arecibo, *J. Geophys. Res.*, *82*, 3243, 1977.

Hays, P.B., T.L. Killeen and B.C. Kennedy, The Fabry-Perot interferometer on dynamics explorer, *Sp. Sci. Instrum.*, *5*, 395, 1981.

Heelis, R.A., Studies of ionospheric plasma and electrodynamics and their application to ionosphere-magnetosphere coupling, *Rev. Geophys.*, *26*, 317, 1988.

Heelis, R.A., J.C. Foster, O. de la Beaujardiere and J. Holt, Multistation measurements of high-latitude ionospheric convection, *J. Geophys. Res.*, *88*, 121, 1983.

Heelis, R.A., W.B. Hanson, C.R. Lippincott, D.R. Zuccaro, L.H. Harmon, B.J. Holt, L.E. Doherty and R.A. Power, The ion drift meter for dynamics Explorer-B, *Sp. Sci. Instrum.*, *5*, 511, 1981.

Heppner, J.P. and N.C. Maynard, Empirical high-latitude electric field models, *J. Geophys. Res.*, *92*, 4467, 1987.

Hines, C.O. and Colleagues, The Upper Atmosphere in Motion, *Geophys. Mono.*, *18*, Am. Geophys. Union, Wash. D.C., 1974.

Hoegy, W.R., S.A. Curtis and L.H. Brace, Electron density irregularities observed on DE-2, in *Effects on the Ionosphere on C^3I Systems*, eds. J.M. Goodman, F.D.

Clorde, J.A. Klobuchar and H. Sorcher, Nat'l. Tech. Info. Serv., U.S. Dept. Comm., Springfield, VA 22167, 1984.

Hoegy, W.R., S.A. Curtis, L.H. Brace, N.C. Maynard and R.A. Heelis, DE observation of equatorial spread-F; evidence for drift waves, *Geophys. Res. Lett, 9*, 995, 1982.

Hoffman, J.H., W.B. Hanson, C.R. Lippincott and E.E. Ferguson, The magnetic ion-mass spectrometer on atmospheric explorer, ***Rad. Sci.**, 8*, 315, 1973.

Hoffman, R.A., M. Sugiura, N.C. Maynard, R.M. Candy, J.D. Cravens and L.A. Frank, Electrodynamic patterns in the polar region during periods of extreme magnetic quiescence, *J. Geophys. Res., 93*, 14515, 1988.

Holmes, J.C. and E.P. Szuszczewicz, Plasma probe system with automatic sweep adjustment, *Rev. Sci. Inst., 52*, 377, 1981.

Holzworth, R.H., M.C. Kelley, C.L. Siefring, L.C. Hale and J.D. Mitchell, Electrical measurements in the atmosphere and ionosphere over an active thunderstorm: 2) Direct current electric fields and conductivity, *J. Geophys. Res., 90*, 9824, 1985.

Huba, J.D. and S.L. Ossakow, Influence of magnetic shear on the current convective instability in the diffuse aurora, *J. Geophys. Res., 85*, 6874, 1980.

Hunton, D.E., E. Trzcinski, L. Wlodyka, G. Federico and J. Dorian, Quadrupole ion/neutral mass spectrometer for space shuttle application, *AFGL-TR-86-0084*, 7 April 1986.

Johnson, D.W. and A. Kavadas, A rocket-borne electric field probe, *Can. J. Physics, 41*, 1980, 1963.

Johnstone, A., A. Coates, S. Kellock, B. Wilken, K. Jockers, W. Rosenbauer, W. Studemann, W. Weiss, V. Formisano, E. Amata, R. Cerulli-Irelli, M. Dobrowolny, R. Terenzi, A. Egidi, H. Borg, B. Hultquist, J. Winningham, C. Gurgiolo, D. Bryant, T. Edwards, W. Feldman, M. Thomsen, M.K. Wallis, L. Biermann, H. Schmidt, R. Lust, G. Haerendel and G. Paschmann, *Nature, 321*, 344, 1986.

Johnstone, A.D., S.J. Kellock, A.J. Coates, M.F. Smith, T. Booker and J.D. Winningham, *IEEE Trans. Nucl. Sci. NS-32*, 139, 1985.

Kelley, M.C. and C.W. Carlson, Observations of intense velocity shear and associated electrostatic waves and turbulence near an auroral arc, *J. Geophys. Res., 82*, 2343, 1977.

Kelley, M.C. and G.D. Earle, Upper hybrid and Langmuir turbulence in the auroral E-region, *J. Geophys. Res., 93*, 1993, 1988.

Kelley, M.C., R.C. Livingston, C.L. Rino and R.T. Tsunada, The vertical wavenumber spectrum of topside equatorial spread-F: estimates of backscatter levels and implications for a united theory, *J. Geophys. Res., 87*, 5217, 1982.

Kelley, M.C., C.E. Seyler and S. Zargham, Collisional interchange instability: 2) A comparison of the numerical simulations with the in-situ experimental data, *J. Geophys. Res., 92*, 10089, 1987.

Kellogg, P.J., S.J. Monson, W. Bernstein and B.A. Whalen, Observations of waves generated by electron beams in the ionosphere, *J. Geophys. Res., 91*, 12065, 1986.

Keskinen, M.J. and S.L. Ossakow, Nonlinear evolution of plasma enhancements in the auroral ionosphere, *J. Geophys. Res. 87*, 144, 1982.

Killeen, T.L., J.D. Craven, L.A. Frank, J.J. Ponthieu, N.W. Spencer, R.A. Heelis, L.H. Brace, R.G. Roble, P.B. Hays and G.R. Carignan, On the relationship between dynamics of the polar thermosphere and morphology of the aurora: global-scale observations from dynamics Explorers 1 and 2, *J. Geophys. Res., 93*, 2675, 1988.

Killeen, T.L., P.B. Hays, Doppler line profile analysis for a multi-channel Fabry-Perot interferometer, *Appl. Opt., 23*, 612, 1982.

Killeen, T.L., P.B. Hays, G.R. Carignan, R.A. Heelis, W.B. Hanson, N.W. Spencer and L.H. Brace, Ion-neutral coupling in the high-latitude F-region: evaluation of heating terms from dynamics Explorer 2, *J. Geophys. Res., 89*, 7495, 1984.

Kozura, J.U., L.H. Brace, T.E. Cravens and A.F. Nagy, A statistical study of the subauroral electron temperature enhancement using DE-2 Langmuir probe observations, *J. Geophys. Res., 91*, 270, 1986.

Krankowsky, D., P. Eberhardt, J.J. Berthelier, J.H. Hoffman, et al., The Giotto neutral mass spectrometer, *Proc. Int'l. Meeting on the Giotto Mission*, ESA SP-169, 1981.

LaBelle, J., M.C. Kelley, and C.E. Seyler, An analysis of the role of drift waves in equatorial spread-F, *J. Geophys. Res., 91*, 5513, 1986.

Lanchester, B.S., H. Rishbeth, T. Nygren, L. Jalonen and T. Turunen, Wave activity, F_1-layer disturbance and a sporadic-E layer over EISCAT, *J. Atm. and Terr. Phys., 51*, 179, 1989.

Leighton A.I., A.H. Shapley and E.K. Smith, The occurrence of sporadic-E, in *Ionospheric Sporadic-E, 2* 166, 1961.

Mathews J.D., Some aspects of metallic ion chemistry and dynamics in the mesosphere and thermosphere, in *Middle Atmosphere Programs, MAP, 25*, 228, 1988.

Mathews J.D. and F.S. Bekeny, Upper atmospheric tides and the vertical motion of ionospheric sporadic layers at Arecibo, *J. Geophys. Res., 84*, 2743, 1979.

Matsushita S. and E.K. Smith (Co-Chairmen), Third Seminar on the cause and structure of temperature latitude sporadic-E, (Logan Utah; Sept. 1971), in special issue *Radio Sci., 7*, 345, 1972.

Matsushita S. and E.K. Smith (Guest Editors), Conference on recent advances in the physics and chemistry of the E-region, (Boulder, Colorado; Aug. 1974), in special issue *Radio Sci, 10*, 229, 1975.

Maynard N.C., T.L. Aggson, F.A. Herrero and M.C. Liebrecht, Average low-latitude meridional electric fields from DE-2 during solar maximum, *J. Geophys. Res., 93*, 4021, 1988.

McKinley D.W.R., *Meteor science and engineering*, McGraw-Hill, New York, 1961.

Mel'nikov V.V., I.A. Savenko, B.I. Savin and P.I. Shavrin, *Geomag. Aeron. 5*, 107, 1965.

Moore T.E., C.J. Pollock, R.L. Arnoldy and P.M. Kintner, Preferential O+ heating in the topside ionosphere, *Geophys. Res. Lett., 13*, 901, 1986.

Mott-Smith, H. and I. Langmuir, The theory of collectors in gaseous discharges, *Phys. Rev., 28*, 727, 1926.

Mozer, F.S., Analyses of techniques for measuring DC and AC electric fields in the magnetosphere, *Space Sci. Rev., 14*, 272, 1973.

Mozer, F.S. and R. Serlin, Magnetospheric electric field measurements with balloons, *J. Geophys. Res., 74*, 4739, 1969.

Mukai T. and W. Miyake, *Rev. Sci. Instrum., 57*, 49, 1986.

Paschmann G., H. Loidl, P. Obermayer, M. Ertl, B. Laborenz, N. Sckopke, W. Baumjohann, C.W. Carlson and D.W. Curtis, The plasma instrument for AMPTE/IRM, *IEEE Trans. Geosci. Remote Sensing, GE-23*, 262, 1985.

Paul W., H.P. Rheinhard and U. von Zahn, Das elektrische massenfilter als massenpektrometer und isotopentrenner, *Z. Physik, 152*, 143-182, 1958.

Paul W. and H. Steinwedel, *Z. Naturforsch, 8a*, 448, 1953.

Paul W. and M. Raether, *Z. Physik, 140*, 262, 1955.

Pedersen, A., U. Fahleson and C.-G. Fälthammar, Determination of the ionospheric density and temperature using a double probe electric field detector, *Space Research XII*, Akademie-Verlag, Berlin, 1369, 1972.

Pedersen, A., R. Grard, K. Knott, D. Jones, A. Gonfalone and U. Fahleson, Measurements of quasi-static electric fields between 3 and 7 earth radii on GEOS-1, *Space Sci. Rev.,22,* 333, 1978.

Pfaff, R.F., Rocket studies of plasma turbulence in the equatorial and auroral electrojets, Doctoral thesis, Cornell University, Ithaca, New York, 1986.

Pfaff, R.F., M.C. Kelley, B.G. Fejer, E. Kudecki, C.W. Carlson, A. Pedersen, and B. Hausler, Electric field and plasma density measurements in the auroral electrojet, *J. Geophys. Res., 89*, 236, 1984.

Pfaff, R.F., M.C. Kelley, B.G. Fejer, N.C. Maynard, and K.D. Baker, In-situ measurements of wave electric fields in the equatorial electrojet, *Geophys. Res. Lett., 9*, 688, 1982.

Phelps A.D.R. and R.C. Sagalyn, Plasma density irregularities in the high-latitude topside ionosphere, *J. Geophys. Res., 81*, 515, 1976.

Reme H., J.A. Sauvaud, C. d'Uston, F. Coton, A. Cros, K.A. Anderson, C.W. Carlson, D.W. Curtis, R.P. Lin, D.A. Mendis, A. Korth and A.K. Richter, Comet Halley-Solar wind interaction from electron measurements aboard Giotto, *Nature, 321*, 349, 1986.

Richmond, A. D., M. Blanc, B. A. Emery, R. H. Wand, B. G. Fejer, R. F. Woodman, S. Ganguly, P. Amayenc, R. A. Behnke, C. Calderon and J. V. Evans, An empirical model of quiet-day ionosphere electric fields at middle- and low-latitudes, *J. Geophys. Res., 85*, 4658, 1980.

Rino, C.L., R.C. Livingston, and S.J. Matthews, Evidence for sheetlike auroral ionospheric irregularities, *Geophys. Res. Lett., 5*, 1039, 1978.

Rodriguez P. and E.P. Szuszczewicz, High-latitude irregularities in the lower F-region: intensity and scale-size distributions, *J. Geophys. Res., 89*, 9575, 1984.

Sablik M.J., D. Golimowski, J.R. Sharber and J.D. Winningham, Computer simulation of a 360° field-of-view "top-hat" electrostatic analyzer, *Rev. Sci. Instr., 59*, 146, 1988.

Scarf F. and Task Group on Solar Space Science (NAS/SSB), Space science in the twenty-first century: imperatives for the decades 1995 to 2015, National Academy Press, Washington, DC, 1988.

Schunk R.W. and E.P. Szuszczewicz, First-principle and empirical modeling of the global-scale ionosphere, *Ann. Geophysicae, 6*, 19, 1988.

Sharber J.R., J.D. Winningham, J.R. Scherrer, M.J. Sablik, C.A. Bargainer, P.A. Jensen, B.J. Mesh, N. Eaker and J.C. Baird, Design, construction and laterating calibration of the angle-researching energy analyzer (AREA): a "top-hat" instrument for auroral research, *IEEE Trans. on Geoscience and Rem. Sensing, 26*, 1988.

Shen J.S., W.E. Swartz and D.T. Farley, Ionization layers in the nighttime E-region valley above Arecibo, *J. Geophys. Res., 81*, 5517, 1976.

Siefring, C.L., Upward propagating electric fields from thunderstorms and VLF transmitters, Doctoral thesis, Cornell University, Ithaca, New York, 1987.

Singh, M., P. Rodriguez, and E.P. Szuszczewicz, Spectral classification of medium-scale high-latitude F-region plasma density irregularities, *J. Geophys. Res., 90*, 6525, 1985.

Singh, M., P. Rodriguez, D.N. Walker and C.S. Liu, A review of ionospheric irregularities at equatorial and high latitudes as observed by the STP/S3-4 satellite, in *The Effect of the Ionosphere on Communication, Navigation and Surveillance Systems*, J.M. Goodman, J.A. Klobuchar, R.G. Joiner, and A. Soicher, eds., U.S. Gov't. Printing Office, 539-548, 1987.

Singh, M., E.P. Szuszczewicz, Composite equatorial spread-F wavenumber spectrum from medium to short wavelength, *J. Geophys. Res., 89*, 2313, 1984.

Smith, E.K., Jr., Worldwide occurrence of sporadic-E, *National Bureau of Standards Circular, 582* U.S. Gov't Printing Office, 1957.

Smith, E.K., Jr., The occurrence of sporadic-E, in *Ionospheric Sporadic-E, 2*, 3, 1961.

Smith, E.K., Jr., World maps of Sporadic-E ($foEs > MHz$) for use in prediction of VHF oblique-incidence propagation, *OT Special Publication 76-10*, U.S. Dept. Comm/Office of Telecommunication, 1976.

Smith, L.G., A sequence of rocket observations of nighttime sporadic-E, *J. Atm. Phys.*, *32*, 1247, 1970.

Smith, Z.K. and J.R. Day, *Rev. Sci. Instrum.*, *42* 968, 1971.

Spencer, N.W., L.E. Wharton, H.B. Niemann, A.E. Hedin, G.R. Carignan and J.C. Maurer, The Dynamics Explorer wind and temperature spectrometer, *Space Sci. Instrum.*, *5*, 417, 1981.

Spencer, N.W., L.E. Wharton, G.R. Carignan and J.C. Maurer, Thermosphere zonal winds, vertical motions and temperature as measured from Dynamics Explorer, *Geophys. Res. Lett.*, *9*, 953, 1982.

Strobel, D.F., Physics and chemistry of the E-region: A review, *Radio Sci.*, *9*, 159, 1974.

Szuszczewicz, E.P., Area influences and floating potentials in Langmuir probe measurements, *J. Appl. Phys.*, *43*, 874, 1972.

Szuszczewicz, E.P., Laboratory simulations of controlled energetic electron-beam-plasma interactions in space, *AIAA Journal*, *21*, 1374, 1983.

3. Szuszczewicz, E.P., Theoretical and experimental aspects of ionospheric structure: a global perspective on dynamics and irregularities, *Radio Science*, *21*, 351, 1986.

Szuszczewicz, E.P., B. Fejer, E. Roelof, R. Schunk, R. Wolf, R. Leitinger, M. Abdu, B.M. Reddy, J. Joselyn, P. Wilkinson and R. Woodman, SUNDIAL: A world-wide-study of interactive ionospheric processes and their roles in the transfer of energy and mass in the Sun-Earth system, *Ann. Geophysicae*, *6*, 3, 1988.

Szuszczewicz, E.P. and J.C. Holmes, Surface contamination of active electrodes in plasmas: distortions of conventional Langmuir probe measurements, *J. Appl. Phys.*, *46*, 5134, 1975a.

Szuszczewicz, E.P. and J.C. Holmes, A versatile plasma probe, *Rev. Sci. Instr.*, *46*, 592, 1975b.

18. Szuszczewicz, E. P. and J. C. Holmes, Observation of electron temperature graduate in mid-latitude Es layers, *J. Geophys. Res.*, *82*, 5073, 1977.

Szuszczewicz, E.P., J.C. Holmes and M. Singh, The S3-4 ionospheric satellite experiment: probe detection of multi-ion component plasmas and associated effects on instability processes, *Astrophys. Sp. Sciences*, *86*, 235, 1982.

Szuszczewicz, E.P., M. Singh and J.C. Holmes, Satellite and rocket observations of equatorial spread-F irregularities; a two-dimensional model, *J. Atm. Terr. Phys.*, *43*, 779, 1981.

Szuszczewicz, E. P. and P. Z. Takacs, Magnetosheath effects on cylindrical Langmuir probes, *Phys. Fluid 22*, 2424, 1979.

Szuszczewicz, E.P., D.N. Walker, H.R. Anderson, and K. Papadopoulos, Suprathermal electron tail distribution in a space-simulation beam-plasma-discharge, in *Active Experiments in Space* (G. Haerendel, ed.) ESA SP-195, ESA Scientific and Technical Publications Branch, STEC, Noordwijk, Netherlands, 189, 1983.

Szuszczewicz, E.P., D.N. Walker, J.C. Holmes and H. Leinbach, Plasma diffusion in a space-simulation beam-plasma-discharge, *Geophys. Res. Lett.*, *6*, 201, 1979.

Temerin, M., Doppler shift effects on double-probe-measured electric field power spectra, *J. Geophys. Res.*, *84*, 5929, 1979.

Tong, Y., J. D., Mathews and W. P. Ying, An upper E-region quarterdiurnal tide at Arecibo?, *J. Geophys. Res.*, *93*, 10,047, 1988.

Wharton, L.E., N.W. Spencer and H.G. Mayr, The earth's thermospheric superrotation from Dynamics Explorer 2, *Geophys. Res. Lett.*, *11* 531, 1984.

Whitehead, J.D., Production and prediction of sporadic-E, *Rev. Geophys. and Space Sci.*, *8*, 65, 1970.

Chapter 3

PROSPECTS OF TOPSIDE SOUNDING

S. A. Pulinets
Institute of Terrestrial Magnetism, Ionosphere and Radiowave Propagation
USSR Academy of Sciences, IZMIRAN
Troitsk, Moscow Region 142092, USSR

1. Introduction

Since the first swallow flight of the Canadian "Alouette", more than 25 years ago, many satellites with onboard topside sounders produced by different countries have been put in orbit. From the source of discoveries in the 1960s, the topside sounders turned into hard workers like their ground-based brothers.

The interpretation and processing of topside ionograms reached the same level of automation as the bottomside ones [Reinisch and Huang, 1982; Huang and Reinisch, 1982]. But nevertheless topside sounding did not shift from the experimental category to routine applications. The gap between topside sounders appearing in orbit even under the efforts of different countries consists of several years. There still exists no national or international system of global monitoring of the ionosphere by the topside sounding technique despite the convincing results of previous satellites. The lack of ionospheric information for the large territories of the ocean and land is one more weighty argument for the creation of a topside sounding global monitoring system. This is one point the present paper will discuss.

The next points of interest are the new techniques in topside sounding such as trans-ionospheric sounding, using stimulated resonances, and natural and artificial noises registered on topside ionograms for monitoring ionospheric parameters.

One can find comprehensive information on the traditional methods of topside sounding in the IEEE Special Issue on Topside Sounding and the Ionosphere [1969], hence in this paper emphasis will be made on nontraditional techniques and on the problems of global monitoring of the ionosphere.

2. The Family of Topside Sounders

The name "topside sounder" formally can be given to more than ten instruments launched by different countries onboard satellites Alouette 1, (1962), Explorer (1964), Alouette-2 (1965), ISIS-I (1969), Cosmos-381 (1970), ISIS-B (1964), ISS-A (1976), ISS-B (1978), EXOS-B (1978), Interkosmos-19 (1979), EXOS-C (1984), Cosmos 1809 (1986), ISEE-1 and 2 (1978), but in fact they could be divided into several groups according to inherent ideas and parameters. Such sounders like Explorer or Cosmos-381 working on several fixed frequencies only proved the possibility of pulsed radio-sounding from onboard the satellite, but they have nothing to do with global monitoring of electron density in large frequency and height domains.

The second group of satellites like EXOS-B, EXOS-C [Oya et al., 1985] or ISEE 1 and 2 [Gurnett et al., 1978] could be named only "relaxation sounders" from the point of view of their tasks and organizing of their work onboard the satellites. They did have a possibility to receive signals reflected from the ionosphere and even sometimes the whole ionogram but their main purpose was stimulation of plasma waves. Their sounding was not systematic from the geophysical point of view.

The rest of the sounders could be divided into three groups according to their technical parameters and contribution to ionospheric science.

2.1 Alouette and ISIS group

The first one is the US-Canadian group including the ISIS-B sounder working from 1974 up to now. The first discoveries concerning the equatorial anomaly, main trough, diurnal, seasonal and other kinds of upper ionosphere variability were made first onboard these satellites. This group of sounders is well described by Franklin and Maclean [1969]. Their technical parameters are presented in Table 1. I'd like to remind you of the main features inherent to these sounders which essentially differ them from other representatives of the family. The fundamental one is their substantially analogous character. Hence the frequency step or resolution is determined by the relation between the sweep rate and pulse repetition frequency. For this reason the fixed grid of frequencies does not exist in contrast to digital sounders. Due to the very small sweep rate, especially in the low frequency range, the frequency step (8.35 kHz) is more narrow than the overall bandwidth of the receiver (40 kHz). This feature made it possible to study in detail the stimulated plasma resonances [Benson, 1982], to observe the sidebands of emitted pulses [Warnock, 1969], to detect the signals emitted during the previous frequency step when the receiver tuned already to the next frequency step [Muldrew, 1969; Vidmar and Crawford, 1985], to study in detail the low frequency natural noises such as AKR [Benson and Calvert, 1979; Benson, 1985; James, 1980].

One interesting feature of the sounder regimes was the special "mixed" mode of operation when the sounder transmitter operated on fixed frequencies during frequency sweep of the receiver [Benson, 1982]. In combination with switching of emitted power it made possible to study nonlinear wave processes and their thresholds. The dubious was selection of orbit heights. Apogee for Alouette-2 was 2980 km and for ISIS-l the apogee was 3500 km. The oblique and ducted traces appearing on the ionograms [Muldrew, 1969] introduced unnecessary complexity in ionogram interpretation and scaling. The main shortcoming of this group of projects was the absence of high capacity memory. This is the principal limitation for global mapping tasks. Nevertheless, they were the first, and most of the results obtained did not lose the actuality and became a cornerstone for modern knowledge of the ionosphere.

2.2 ISS-B sounder

The next representative of the ionosonde family was the topside sounder launched from the second attempt in February 1978 onboard the Japanese satellite ISS-B [Nishizaki et al., 1982]. The principal parameters of the sounder are shown in Table 2. The main difference from ISIS-like satellites is the digital-generated grid of frequencies with a step of 100 kHz. From the point of view of resolution it is a more rough device in frequency domain and apparent range domain, but in this project a big advance was made from the point of view of global mapping [Atlas of Ionospheric Critical Frequency, 1979] and onboard processing of sounder information. The new nontraditional techniques were used for monitoring ionospheric parameters. These techniques will be discussed in a separate paragraph.

As concerns mapping, the main means was the onboard tape recorder with the recording time of 115 min, which is nearly equal to satellite orbit period (107 min). In the mode TOP-B when the topside ionograms were obtained, the sequence of measuring was 64 s with ionogram duration of 16 s. So during the recording time nearly 100 ionograms were stored, representing a whole revolution around the Earth. Every day 3-4 such orbits were transmitted to the ground-based telemetry station. Taking into account the shift of the satellite orbit in local time, 12 min per day, nearly four months were necessary to build the set of UT maps (see Figure 1), but the problem of mapping will be discussed in section 3. A result of the project was a series of summary plots and maps for the different periods of the satellite activity published by the Radio Research Laboratory of the Japan Ministry of Posts and Telecommunications [Summary Plots, 1983]. It was the main contribution of the project to ionospheric science.

Table 1

SUMMARY OF SYSTEM PARAMETERS FOR THE ALOUETTE AND ISIS TOPSIDE SOUNDERS

	Alouette 1	Alouette II	ISIS-I	ISIS-B
Sweep range (MHz)	0.5 - 12	0.12 - 14.5	0.1 - 10 0.1 - 20	0.1 - 10 0.1 - 20
Sweep rate (MHz/s)	≃ 1	≃ 0.13 below 2 MHz ≃ 1 above 2 MHz	0.31(0.1-2 MHz) 0.88(2-5 MHz)	0.375(0.1-2MHz) 1.125(2-5 MHz)
Frame time (s)	18	32	19 or 29	14 or 21
Fixed frequencies (MHz)	-	-	0.25 0.48 1.00 1.95 4.00 9.303 0.82*	0.25 0.48 1.00 1.95 4.00 9.303
PRF(Hz)	63	30	30 and 60	45
Pulse width (μs)	100	100	98	87
Transmitter power (W)	100	300	400	400
Receiver AGC attack time (ms)	12	520 (below 2 MHz) 60 (above 2 MHz)	60	60
Decay time (ms)	46	120 (below 2 MHz) 12 (above 2 MHz)	12	12

*Resonance mode only

Table 2

ISS-B SPECIFICATION PARAMETERS

Frequency range	0.5 - 14.8 MHz
Frequency step	100 kHz
Pulse width	300 μs
Pulse peak power	150 W
Pulse repetition rate	9 Hz
Bandwidth of the receiver	7 kHz
Sampling of echo signals	33.3 km in apparent range

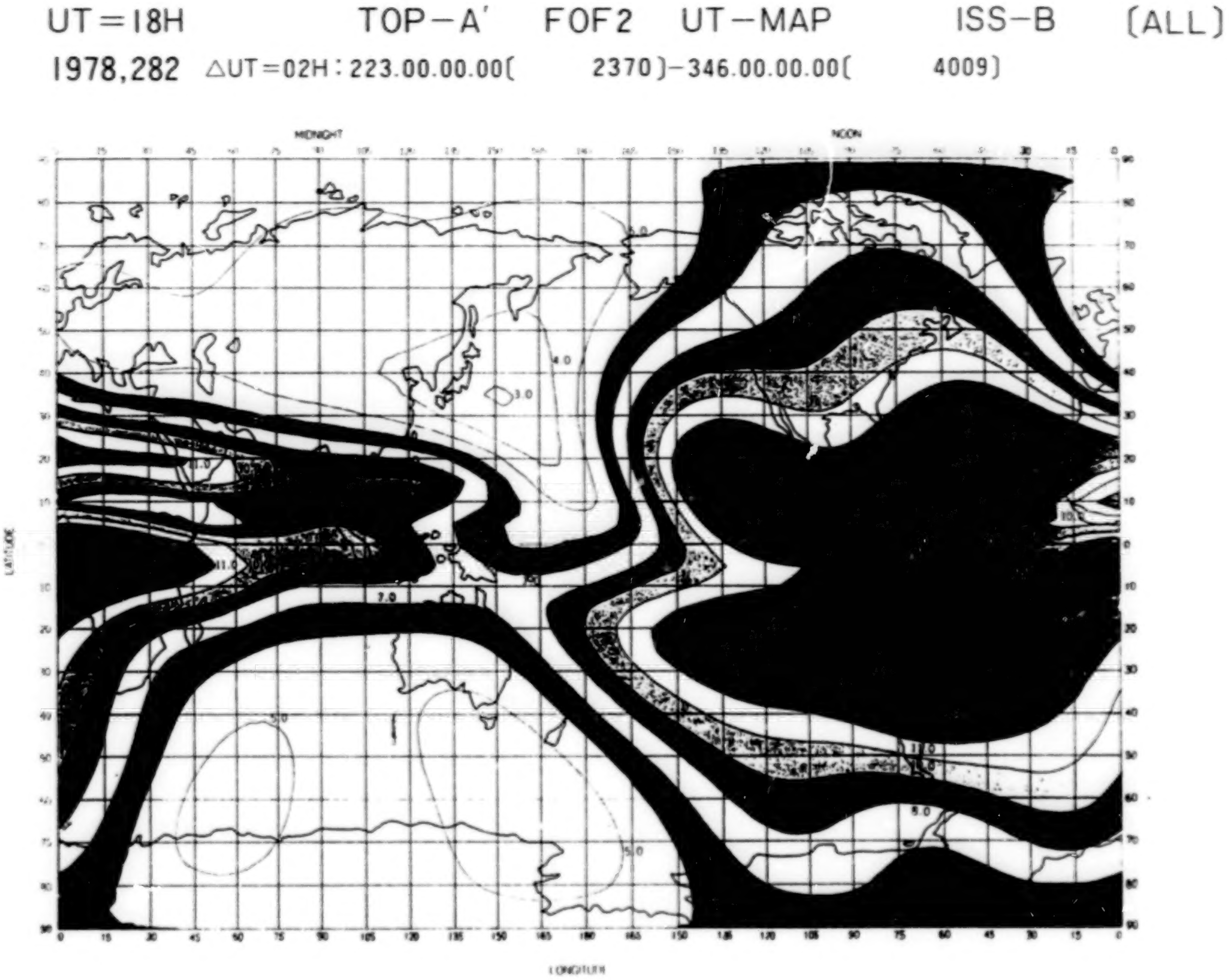

UT=18H TOP-A' FOF2 UT-MAP ISS-B (ALL)
1978,282 △UT=02H : 223.00.00.00(2370)-346.00.00.00(4009)
MIDNIGHT
NOON
LATITUDE
LONGITUDE
4.0
3.0
11.0
7.0
5.0
10.0

Figure 1. Example of ISS-B global UT-map.

It is necessary to mention the new approach to the automation of the topside ionograms processing [Igi and Aikyo, 1986]. It concerns mainly the automatic identification of the resonance spikes but the algorithm (see Figure 2) suggested could be used for onboard processing of the ionograms and automatic identification of O and X traces without polarization measurements.

2.3 IS-338 sounder

Dealing with many ionospheric scientists I heard the regret about the absence of detailed information in English concerning the Soviet topside sounder. To fill this gap I feel a necessity to adduce here a short description of the sounder. The principal designer of this satellite ionospheric station was Dr. G. V. Vasiliyev [Vasiliyev et al., 1980a,b]. It was installed onboard two satellites: Interkosmos-19 (February 1979-March 1982, inclination 74, perigee 500 km, apogee 980 km) and Cosmos-1809 (December 1986 - June 1989, inclination 82.5, circular orbit with height 960-980 km). In principle, the general functional diagram onboard both satellites did not differ except for the new system of ionogram coding on the Cosmos 1809 satellite. The system consisted of the sounder itself, antenna system, coding device and operative memory (onboard Cosmos 1809 -- two coding devices) (Figure 3). The antenna system consisted of two crossed dipoles 50 m tip-to-tip and 15 m tip-to-tip and filters of high and low frequencies. The dipoles were situated parallel to the Earth's surface and under 45° to velocity vector of the satellite (Figure 4). The satellite was 3-axes stabilized. The sounding in the lower frequency band up to 5 MHz was carried out on the 50 m dipole and in the higher frequency range from 5 MHz on the short 15 m dipole. The sounder parameters are displayed in Table 3.

The timeframe and composite video signal are shown in Figure 5. Depending on telemetry memory mode, the sounder is switched on by the sync pulse every 8, 16, 32, or 64 s (Figure 5a) but the ionogram duration is always 6.04 s. Information about the switching of the sounder in the form of 6.04 s pulse is transmitted by the telemetry (Figure 5b) and special pulse is used for control of the onboard devices (Figure 5c). It is necessary to have in mind that some delay of 0.3 - 0.5 s takes place from the switching sync to the beginning of sounding. This time is used for the reaching of stationary state of all electronic circuits after the power supply switches on. Every ionogram starts with the pulsed marker of 0.136 s duration filled by the 15 kHz sinusoidal signal (Figure 5d) which is necessary for ground-based equipment to synchronize the hard copy registration device. This marker is used during transmitting of ionograms by the PM modulation telemetry system. Then the 338 pulses from 0.3 up to 15.95 MHz are emitted with the 17.07 ms period. At the end of the ionogram the pulsed marker is formed, filled by the sinusoidal 10 kHz signal. The detailed structure of the video signal between the neighborhood frequencies is shown in Figure 5e. After the previous frequency signal registration, the "0" level is registered during 1.6 ms, then the negative sync of 533 μs duration is formed for the registration equipment synchronization. At the same time the heterodyne is tuning from one frequency to the next. Then the positive pulse is formed, the duration of which contains information on the natural noise level on the given frequency at the receiver input (the AGC pulse). Then follows the pulse shown by the stroked line which appears on the frequencies multiple to 1 MHz (the frequency mark). Every 5 MHz the frequency mark appears not only on the frequency multiple to 5 MHz, but on the next frequency too (the double mark). Then the negative "zero height" pulse is generated by 13 μs duration and the positive 133 μs duration pulse coincides with the pulse of the sounder transmitter. Then the receiver output waveform is registered contained in the reflections from ionosphere, natural and artificial noises.

2.4 IS-338 ionogram format

The presence of telemetry systems of two types onboard satellites (Figure 2) gave a possibility to vary the type of transmitted information and the duration of sounding periods. The

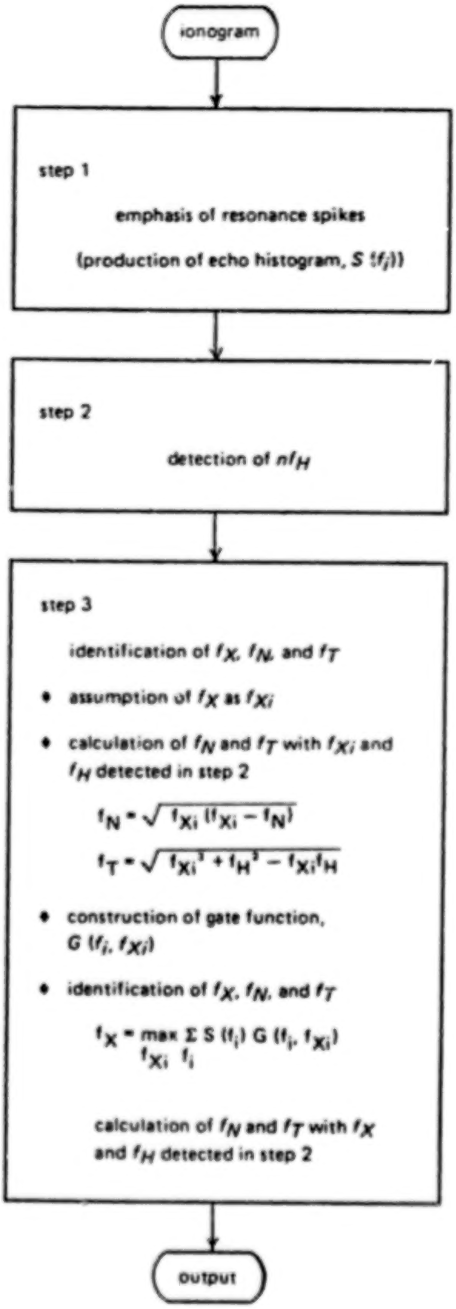

ionogram
step 1
emphasis of resonance spikes
(production of echo histogram, $S(f_i)$)
step 2
detection of nf_H
step 3
identification of f_X, f_N, and f_T
• assumption of f_X as f_{Xi}
• calculation of f_N and f_T with f_{Xi} and f_H detected in step 2
$f_N = \sqrt{f_{Xi}(f_{Xi} - f_N)}$
$f_T = \sqrt{f_{Xi}^2 + f_H^2 - f_{Xi}f_H}$
• construction of gate function, $G(f_i, f_{Xi})$
• identification of f_X, f_N, and f_T
$f_X = \max_{f_{Xi}} \sum_{f_i} S(f_i)\,G(f_i, f_{Xi})$
calculation of f_N and f_T with f_X and f_H detected in step 2
output

Figure 2. Flow chart for automatic resonance spike identification.

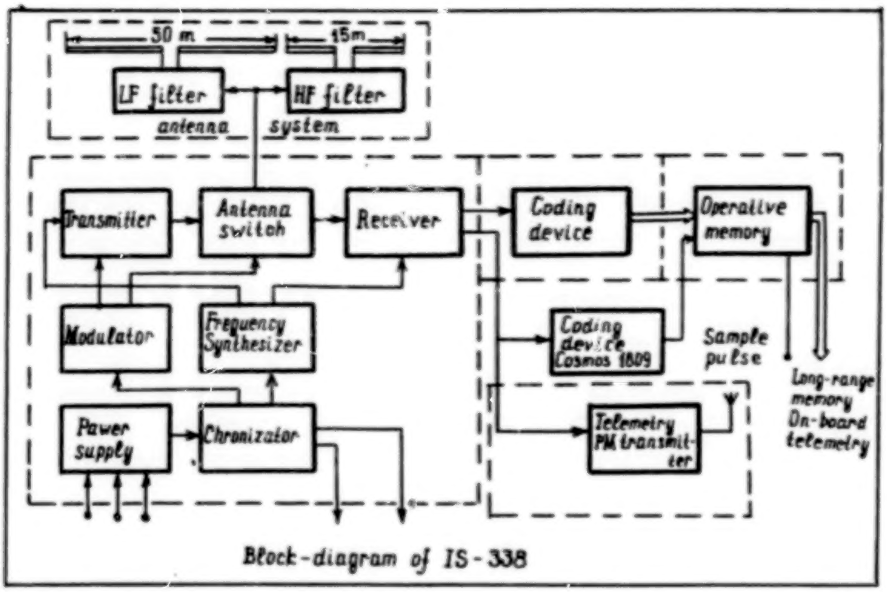


Figure 3. Block diagram of IS-338.

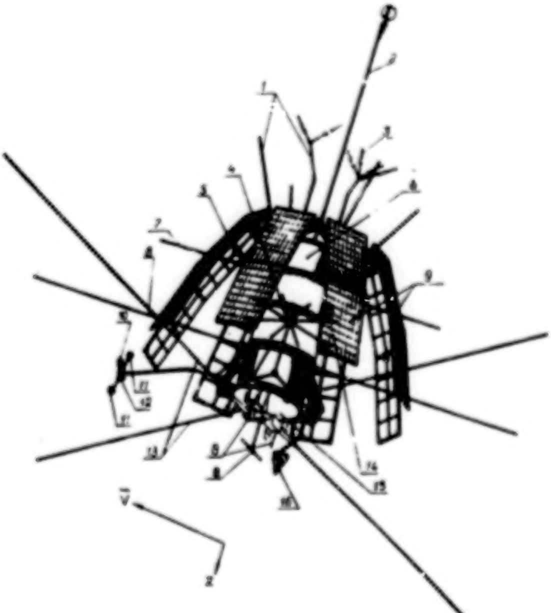


Figure 4. Intercosmos-19 satellite. 1 - command radio line antennas; 2 - gravitation-damping device; 3,4 - antennas of PM and wideband FM telemetry, 5 - P-4 probe; 6 - service equipment module, 7 - radiospectrometer antenna; 9 - solar cell panels; 10 - KM-3 probe; 11 - VLF electric probe; 12 - VLF magnetic probe; 13 - IS-338 antenna system; 14 - soft-particle spectrometer; 15 - module of scientific devices; 16 - high-capacity telemetry system antenna.

ionogram transmitted by analogous PM VHF telemetry system in the 136-138 MHz frequency band is presented in Figure 6a. The full complex video output shown in Figure 5 was transmitted but the quantity of ionograms transmitted and the range of latitudes and longitudes was limited by the direct visibility of the satellite by the ground-based telemetry receiving station. In the direct transmission mode the repetition period of sounding was always equal to 8 s and nearly 70 ionograms in average were registered during one pass of the satellite over the telemetry station.

The digital ionogram format is shown in Figure 6b. Ionograms of this type are obtained with the help of a coding device and operative memory, then they are copied to the onboard telemetry memory. Depending on the ionogram repetition period (8, 16, 32, 64 s) the geographical scope changed. When the repetition period was 64 s, it was possible to put in memory nearly 10 or 11 full revolutions which corresponds to nearly 250° in longitude. In this format only three points were registered on every frequency. These points corresponded to time delays of the first three pulsed signals exceeding the threshold determined by the AGC. The delay time was transformed to 8-bit numbers with the sample frequency of 15 kHz which corresponds to 10 km in apparent range scale. If any signal on a given frequency did not exceed the threshold, the number corresponding to 2000 km delay was generated. The two numbers equal "3" and two numbers equal "6" were the marks of beginning and the end of the ionogram. The volume of the ionogram was 1014 bytes. The details such as plasma resonances were lost during this type of registration but the global survey during the short period of time was the prize.

For the Cosmos-1809 satellite the new type of coding was developed by Dr. M. D. Fligel and Dr. G. V. Vasiliyev. It was named digital-analogous ionogram and is shown in Figure 6c. The additional coding device and operative memory were installed onboard the satellite (Figure 3). As in the previous format, the sampling frequency was 15 kHz, which corresponds to 10 km in apparent range. If the output signal within the 10 km apparent stretch was higher than the threshold level the logical "1" was generated, otherwise the logical "0". So for the one definite frequency the apparent range 0 - 2000 km contained 200 bits of information. These 200 bits were divided into 25 groups of 8 bits -- the 25 quasi-byte numbers stored in the operative intermediate memory. These 25 bytes were supplemented by 3 bytes more, two of which contained information on the 16-level AGC voltage and the third one the frequency mark. As a result, one ionogram contained 338.28 $\simeq$ 9.5 bytes. To mark the beginning and the end of the ionogram nearly 256 numbers were recorded containing only "1". The onboard memory gave a possibility to store nearly 250 ionograms. Taking into account the ionogram repetition period of 32 s and the satellite orbit period of 104 min, the $\simeq$ 1.3 of satellite orbit were stored, which was close to the ISS-B regime with the exception that on the Cosmos-1809 satellite the sampling of ionograms was two times more often.

2.5 Future topside sounders

One of the advanced suggestions was made by Mathwich et al. [1981] where all ideas of ground-based digisondes were included. Up to now this project is not realized. It seems that it would be useful to remind readers the main ideas of the paper.

The fundamental problem of topside sounding is automatic scaling of obtained ionograms to exclude operator intervention. The very important steps in this direction were made in the suggestion. These are the coherent signal processing by special signal coding and tagging of the right-hand circulated mode and left-hand circulated mode by combination of special antenna systems, separate emission of right-hand circulated pulses (RCP) and left-hand circulated pulses (LCP), and using different codes for RCP and LCP. Great attention was paid to the extensive onboard data compression. The second important feature of the suggested mission is the decreasing the power to 30 W of emitted pulses. As was shown by Pulinets and Selegey [1983, 1986] the large-scale modification of near satellite plasma takes place during and after emitting of

Table 3

THE IS-338 TOPSIDE SOUNDER PARAMETERS

Number of frequencies	338
Frequency range	0.3 - 15.95 MHz
Crystal synthesizer range	35.3 - 54.95 MHz
Frequency step	25 kHz in 0.3 - 1.5 MHz range 50 kHz in 1.5 - 15.95 MHz range
Synthesizer switching time	- 2 ms
Emitted pulse duration	133 μs
Repetition rate	58.6 Hz
Ionogram duration	6.04 s
Ionogram repetition period	8, 16, 32 or 64 s (depending on telemetry memory mode)
Apparent height range	0 - 2000 km
Height resolution	10 km
Receiver sensitivity	5-10 μV with S/N=3
Receiver IF band width	12 kHz
Dynamics	80 dB
Supplied power	50 W during sounding 0.2 W during pauses
Pulsed emitted power	200 - 300 W
On every 338 frequencies one pulse was emitted	

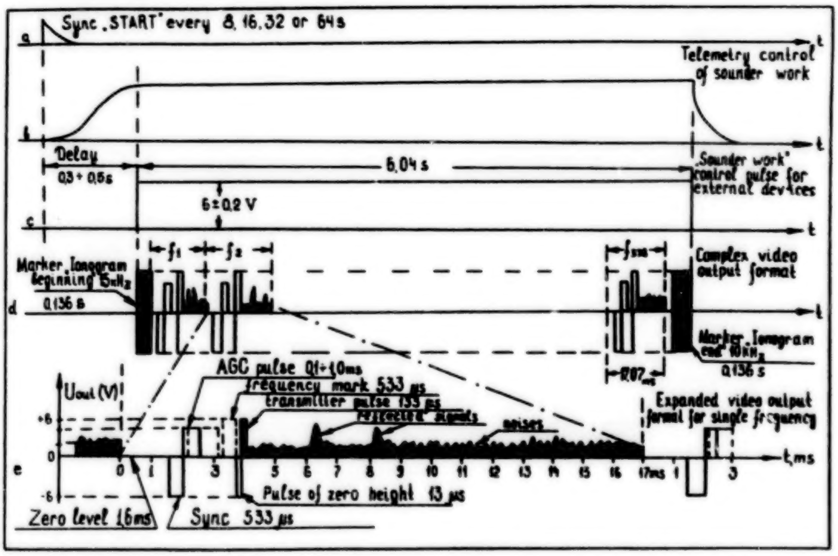

Sync „START" every 8, 16, 32 or 64 s
a
t
Telemetry control of sounder work
b
t
Delay
0,3 ÷ 0,5 s
6,04 s
„Sounder work" control pulse for external devices
6 ± 0,2 V
c
t
f_1
f_2
f_{338}
Marker „Ionogram beginning" 5 kHz
0,136 s
d
Complex video output format
t
Marker „Ionogram end" 10 kHz
0,136 s
17,07 ms
AGC pulse 0,1 ÷ 1,0 ms
frequency mark 533 μs
transmitter pulse 133 μs
reflected signals
noises
U_{out} (V)
+6
0
-6
e
0
1
3
5
6
7
8
9
10
11
12
13
14
15
16
17 ms
1
3
t, ms
Expanded video output format for single frequency
Pulse of zero height 13 μs
Zero level 1,6 ms
Sync 533 μs

Figure 5. Video output and timeframe of the IS-338 sounder.

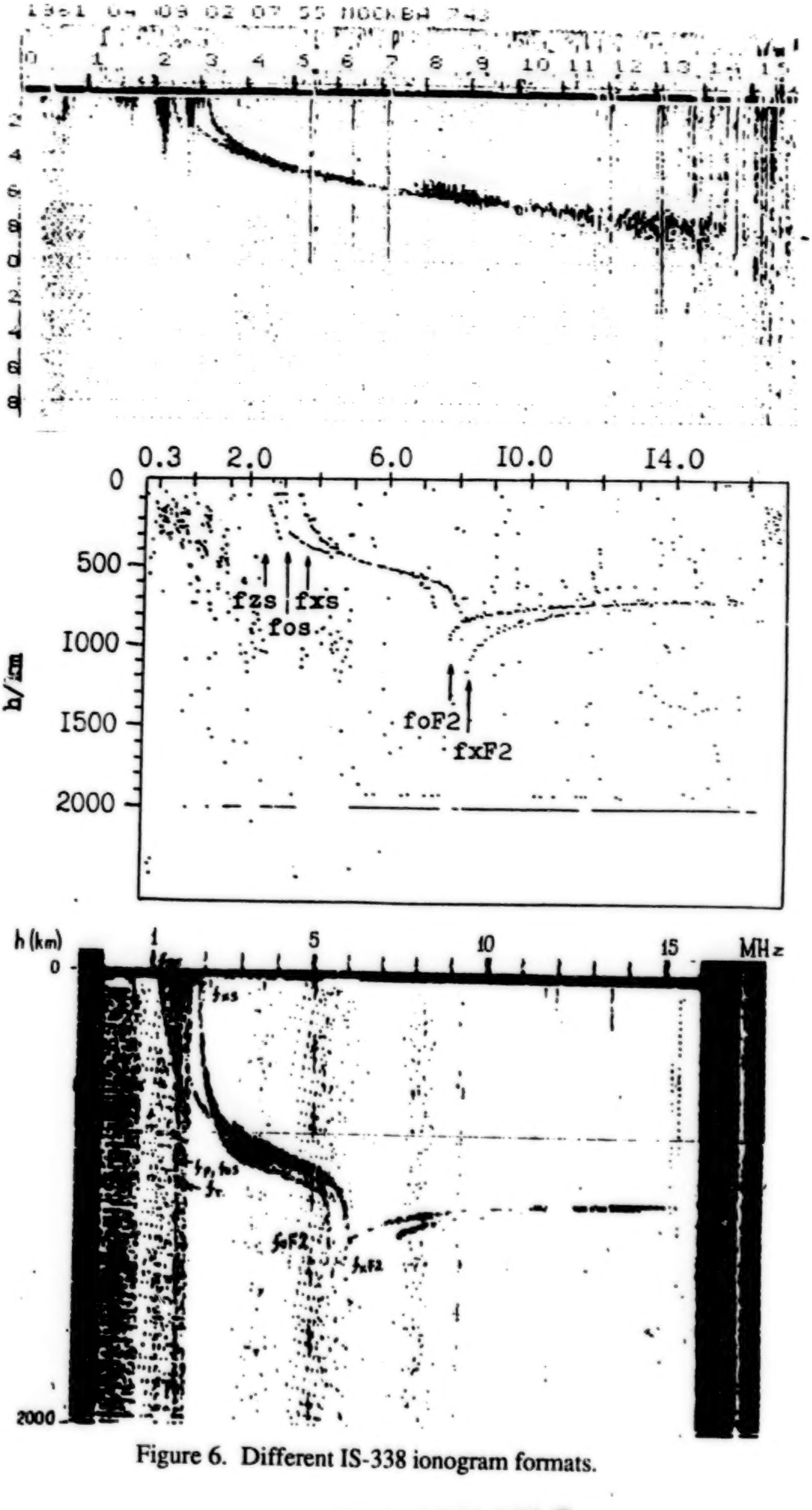


Figure 6. Different IS-338 ionogram formats.

BEST COPY AVAILABLE

powerful RF pulses of the sounder transmitter. It creates problems for different kinds of in situ onboard measurements as from the point of view of real values of measured parameters of undisturbed plasma so from the point of view of simple electronic interference. The diminishing of the emitted power by the order is a great step to the last concept of the mission as a guest mission on any spacecraft having the appropriate orbit. The questionable feature of the mission is the duration of ionogram sampling (to 40 s). In highly dynamical auroral and polar regions it may create problems in ionogram interpretation, and many details of the morphological structures of the high latitude ionosphere might not be resolved.

3. The Global Mapping of the Topside Ionosphere

It is hoped that such projects as WITS and SUNDIAL as future parts of the STEP Project are manifestations of a renaissance in ionospheric science. One of the main purposes of the SUNDIAL project, for example, is to know "the ionospheric state at any place, at any time and under any conditions" [Szuszczewicz et al., 1988]. Global mapping of the ionosphere with the help of satellite topside sounders gives such a possibility. But to not fall in euphoria it is necessary to realize all difficulties and limitations of the technique. Topside sounding is one of the most complex techniques of remote sensing from the points of view of

a - the great variability of ionospheric parameters
b - interpretation of ionograms (propagational problems, oblique layers, ducting, spread echoes, etc.)
c - scaling of ionograms
d - four-dimensional representation of the results of sounding.

3.1 From ionogram to map

The problems of topside ionogram scaling and vertical profile calculations were sufficiently discussed [Hagg et al., 1969; Jackson, 1969; Huang and Reinisch, 1982; Reinisch and Huang, 1982; Serebryakova, 1986; Sotsky, 1986] and it seems it is not necessary to discuss them in this paper. I'd like to discuss only one of the problems of ionogram interpretation: the role of the trace of the Earth's reflection echoes. If we assume vertical propagation and horizontal layered structure of the ionosphere, the critical frequencies scaled from the topside reflection trace (TRT) and the Earth's reflection trace (ERT) have to coincide (Figure 7a). In practice, however, we observed many cases where the traces did not coincide as in Figure 7b (too short) or Figure 7c (overlapping). In the case of Figure 7b we deal with the very thick F layer [Kochenova and Fligel, 1989] which can lead to errors in the determination of the F peak height. The case of Figure 7c is connected with oblique propagation of sounder pulses due to the nonhorizontal position of the F layer [Fligel, 1989]. Figure 8 explains this situation. The height scaled from TRT does not give real height values but h•cos α (α is the angle which the F layer makes with horizontal). On the other hand, the ERT gives the real value of the critical frequency f_{oe} due to the vertical propagation of ERT pulses. But the observed Earth reflections do not reach this value screened by the oblique F layer. The last observed point of ERT is close to $f'_{oe} \simeq f_{oe} \cdot \sec \alpha$. In the cases of Figures 7b and 7c the ERT is a representative trace for the accurate topside profile calculation [Kovalev et al., 1989].

The next step is to combine the obtained set of profiles to represent the global distribution of electron density. Two techniques were used: vertical cross sections of the ionosphere during one satellite revolution over the Earth or the plane distribution of one of the ionospheric parameters (f_0F2 distribution, for example) versus geographical or geomagnetical coordinates. In the first case we have a snapshot of the state of the ionosphere [Vasiliyev et al., 1981]. On Figure 9 the global distribution of electron density is shown for the different latitudes but for some fixed longitude determined by the satellite orbit plane position.

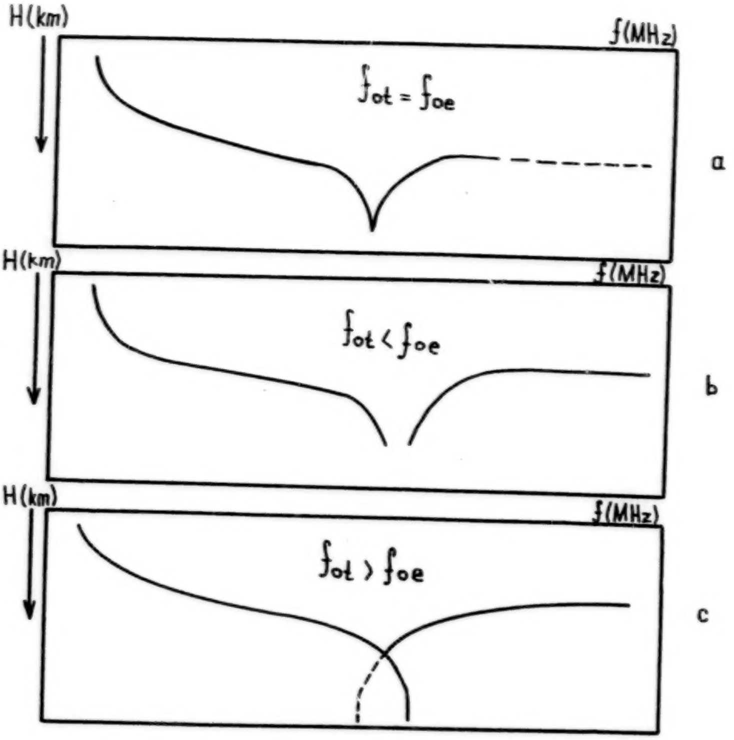


Figure 7. Different cases of topside and ground-reflected traces.

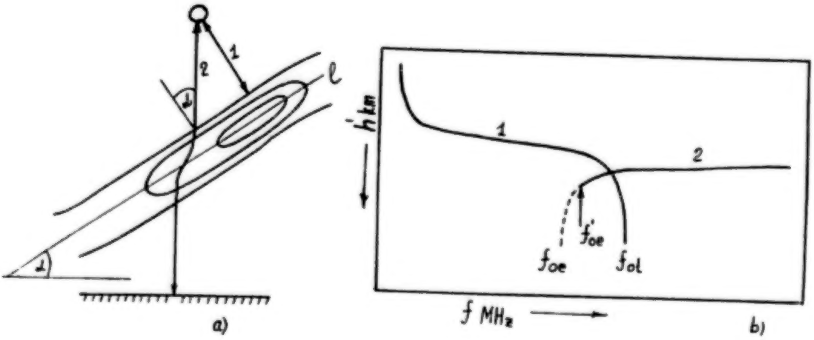


Figure 8. Overlapping on topside ionograms.

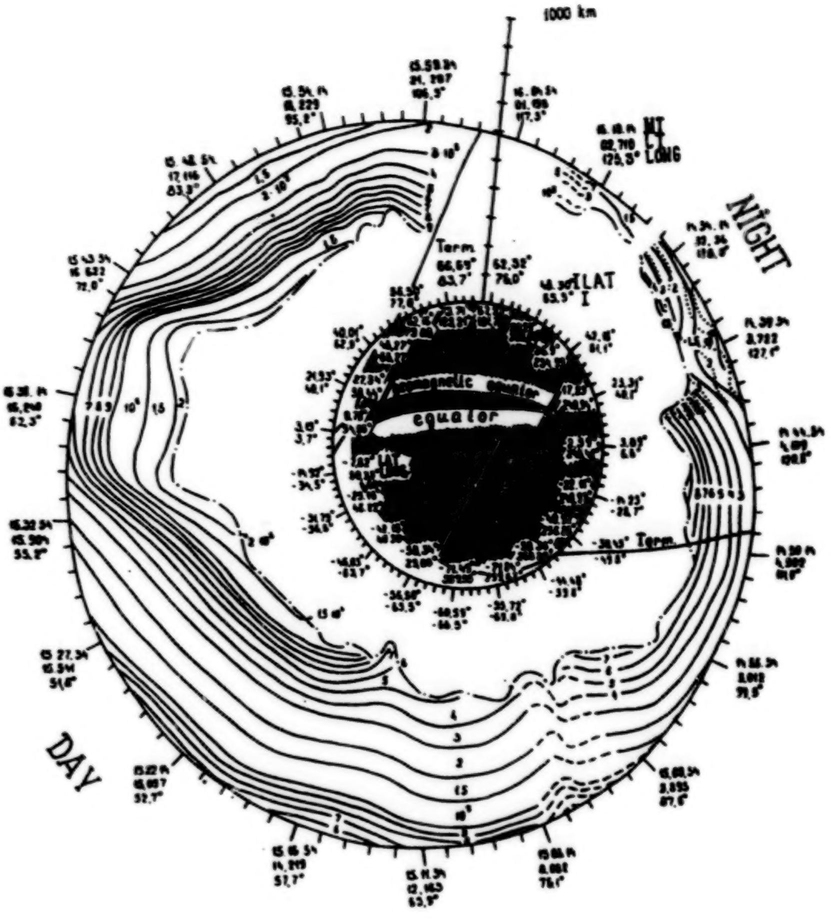


Figure 9. Global distribution of electron density for one satellite orbit.

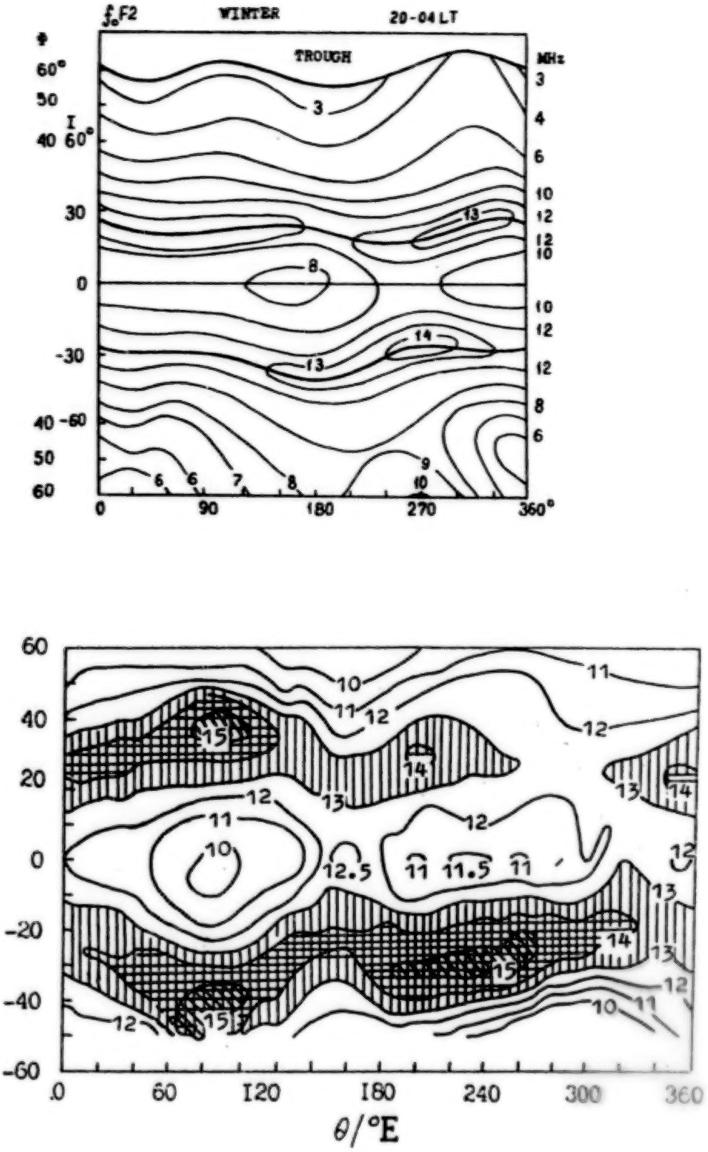


Figure 10. LT-maps for f_oF2 solar maximum (1979-81). a) winter nighttime conditions; b) summer daytime conditions.

The most difficult step is the transition from single orbit data to real maps. The shift in longitude during one full revolution of the satellite with orbital period of order of 100 min, taking into account Earth's rotation velocity, use to be near 25°. So it is necessary to have more than 14 full revolutions to scan over all 360° of longitude which equals 24 hours of continuous sounding. Here one more problem arises: the problem of LT and UT maps. The satellite on the near polar orbit is "tied" to the definite local time. This means that the fixed latitude the satellite crosses at the same local time. The satellite orbit is divided by two parts with 12 hours difference in local time, each part "tied" to its local time. So using information collected during 24 hours it is possible to build two global maps for two fixed meanings of the local time, so-called LT-maps. But one must conceive clearly that an LT-map is in some sense the artificial formation (situation with the same local time all over the world is physically impossible). These maps show how the ionospheric parameters differ at the same local time on different longitudes. It is a very convenient technique to build such maps for studying the global longitudinal effect [Diominov and Karpachov, 1986; Karpachov, 1987; Diominov et al., 1987; Ben'kova et al., 1988]. Two such maps for night and for day conditions are shown in Figure 10.

Another global representation is the so-called UT-map. It represents the real state of the ionosphere at a given moment of UT, its "instant picture". It is evident that having the satellite "tied" to the local time it is impossible in principle to obtain this kind of map for a short period of time. Let us try to estimate how many times it is necessary for a satellite like IK-19 to look over the whole 24 hours of LT and UT. For the IK-19 satellite the diurnal time shift in local time was nearly 12 min. In this case nearly 4 months is necessary for the whole survey in local time. If we take into account the ascending and descending parts of orbit it is possible to double shorten this interval, but in reality a satellite does not work continuously. It is connected with the problems of satellite management: refreshing of onboard memory, sending of steering commands and programs and so on, and all the same the building of UT-maps takes more time than a season. In fact the building up of one UT-map is equal to building up a whole set of LT-maps for 24 hours. The UT-map will reflect the real ionospheric situation only when the period of building up corresponds to the same season. It is clear also that the data must be selected in solar and geomagnetic activity. In the case of solar maximum these conditions to a great extent restrict the range of suitable material. So even for one instant of UT the task of building a global map is very complex. To build a really good U T-map it is necessary sometimes to take data from different years of satellite operation for the same season. We don't speak yet on the equipment restrictions onboard the satellite. For example, the Japanese satellite ISS-B, due to the small volume of onboard memory, could receive information only from 3 - 4 orbits a day. This fact makes it impossible to build adequate UT-maps on the basis of these data.

One of the advantages of IK-19 was a large-memory regime when information was collected from 10 - 11 orbits without a break (one ionogram every 64 s), which constitute 17 hours of sampling and 250 degrees in longitude. In addition, this mode of operation, after a small break, could be repeated. This permitted to build up the LT-map for the short period of time of nearly two days. These maps have some shortcomings connected with the incomplete polar orbit of the satellite. The change in local time along the orbit with 74° inclination within 60° interval was sufficiently large ($\Delta LT \simeq 4$ h). This difference is more noticeable if we work in magnetic coordinates. For example, for fixed geomagnetic latitude 60°Φ in the Southern Hemisphere $\Delta LT \simeq 2$ h. Although the quantity of selected data must be increased before we could prove that this change doesn't matter for our specific problem. We analyzed this question and at least two problems could be solved without narrowing of the LT interval: the strong disturbances in the upper ionosphere and studying of global longitude effect in the midnight ionosphere. It is due to the fact that the observed variations are essentially larger than the temporal variations.

So what to do with the mapping problem? It is evident that in the frame of one satellite the problem of LT-UT maps cannot be solved. Some improvement could be reached only by the continuous transmitting of information to the Earth without putting it into memory. The retranslation geostationary satellite is a means of decision. But the problem of 2 -4 months for UT maps remains. To shorten the period a system of satellites is necessary.

3.2 The proposed satellite system of global monitoring and mapping of the ionosphere

As we realized from the previous paragraph, for proper monitoring and mapping from satellites, two main demands should be fulfilled: continuous sounding and sufficiently fast overlapping of 24 hours of LT time. For the first, different approaches could be used. The simplest one is high capacity onboard memory (nearly 24 hours) like on the Interkosmos-19 satellite. For mapping it is acceptable, but for operative monitoring it is a wrong decision due to the delay in receiving data. The possible solution is to have only one-orbit memory like on the ISS-B satellite and to transmit information after every pass. But it is a difficult task: too much effort is necessary for satellite control, and several receiving stations spread in longitude should be used, due to the fact that one receiving station can receive only four orbits per day, owing to shift of the satellite orbit along longitude. The utilization of a geostationary satellite as a retranslator arises. Several variants are possible. First is a direct continuous transmission from the sounding satellite to the geostationary satellite without any remembering. It is a good suggestion but it needs a sophisticated antenna system to track the retranslator by the moving satellite. The intermediate variant is possible: to transmit to the geostationary retranslator the information remembered every revolution.

The second problem of mapping and monitoring is coverage of different LT to build the real UT-maps. Unfortunately, only an expensive way is possible -- using several satellites which are put on orbits shifted one from another at definite intervals of LT. For the first time this task was set by Mathwich et al. [1981]. It seems that four satellites would be sufficient for any sophisticated task of monitoring, modeling, and so on. The problem is how to arrange them, what local times are decisive for the state of the ionosphere. The different arguments could be put forward: the midnight - noon sector, the dusk - dawn sector, etc. Every period of time has its own influence on the ionosphere. The way out lies, as it seems, not in the selecting of definite LT sectors, but in refusion from strict sun - synchronized polar orbits. The 80 - 85° inclination orbits shift during a day on 8 - 12 minutes. Such shift gives a possibility to cover all LT sectors during approximately two weeks. The schematic representation of the proposed system of ionospheric monitoring is shown in Figure 11.

Naturally, topside sounding does not exist in isolation from other sounding techniques and a systematic approach should be used in ionospheric service. This problem was raised by Danilkin [1987a] where cooperation of different radio means is discussed (see Figure 12). The use of different sources of ionospheric information permits to make more profound reconstruction of the F region [Gulyaeva, 1988]. Use of radio beacons onboard the satellites permits to build computer-aided 3-dimensional tomography pictures of the ionosphere [Austen et al., 1988].

4. Nontraditional Techniques in Topside Sounding

The standard work connected with evaluation of height-frequency-density characteristics of the ionosphere are always accompanied by attempts to extract more information. As an example, the studying of topside plasma resonances could be given [Benson, 1982; Oya, 1970, 1971]. We will try to discuss only such techniques which give a possibility of global monitoring of some ionospheric parameter or improve the topside sounding technique itself.

As a first attempt, Benson's [1972] paper could be considered. The global distribution of different kinds of plasma resonance were studied, many interesting irregularities were discovered,

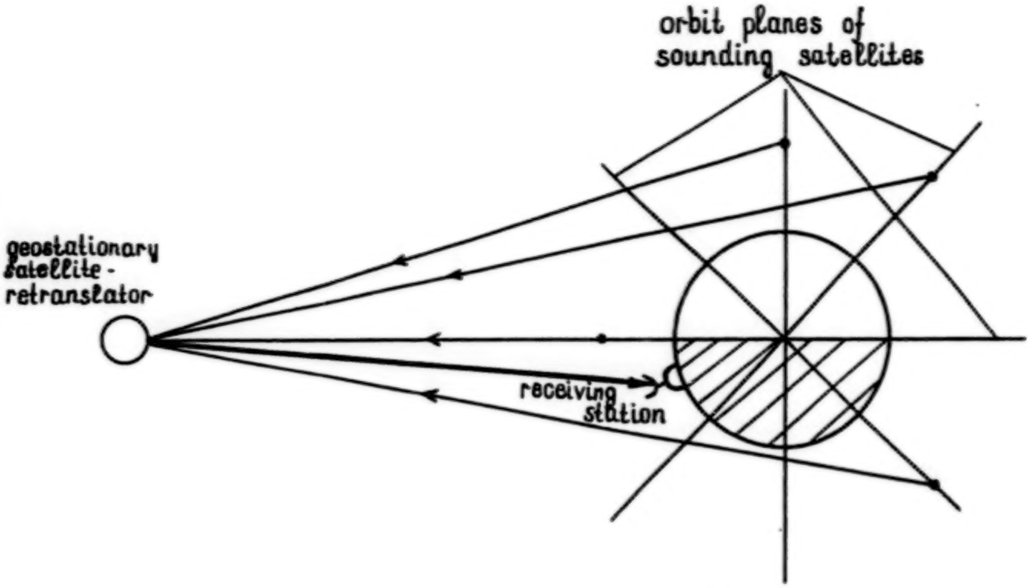


Figure 11. Satellite system for global monitoring of the ionosphere.

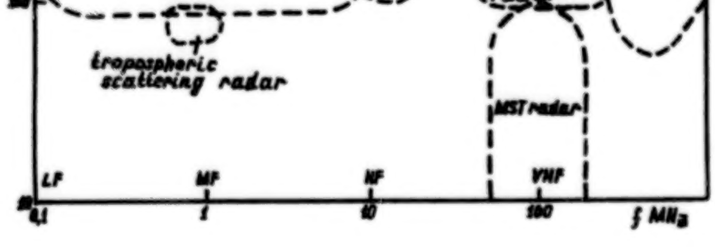


Figure 12. Systematic approach to ionospheric sounding.

but this work has not received further development, primarily due to the problem of resonance scaling. Automatic scaling [Igi and Aikyo, 1986] has not been developed.

4.1 The DNT resonance as an indicator of small-scale irregularities in the ionosphere

As has been shown by Denisenko et al. [1987] and Bratsun et al. [1988], the weakening of O-traces on topside ionograms in the frequency band $F_p < f < f_{UHR}$ is connected with the transformation of ordinary waves into slow extraordinary left-hand polarized Z-waves in the presence of small-scale irregularities. As a manifestation of scattering on irregularities of Z-waves the diffuse resonance appears on the topside ionograms (see Figure 13) named by Benson [1982] DNT resonance. It appears between the local plasma frequency f_p and the frequency of upper hybrid resonance f_{UHR} in the case of $f_p > f_{He}$, and between the local gyrofrequency f_{He} and upper hybrid frequency when $f_p < f_{He}$. The apparent height of the resonance appearing on topside ionograms is a mixed value which is a manifestation of its duration and amplitude simultaneously. This height of DNT resonance scaled from ionograms along the satellite orbit may be used for estimating the scale of irregularities development. Figure 14 represents the distribution of the resonance height vs magnetic latitude for four orbits of Cosmos-1809 satellite for day conditions during summer solstice, May - June 1987 [Pulinets et al., 1989]. It could be used to study a global distribution of small-scale irregularities in the upper ionosphere.

4.2 Natural noises

On Figure 14 one can mark the groups of points in the vicinity of the magnetic equator and on latitudes near +80° which are situated at the top of the figure. They manifest that the DNT resonance height exceeds the whole height range of the ionogram. The special experiment was conducted onboard the "Sibir" icebreaker where the receiving station was installed during its voyage to the North Pole in summer 1987. The task of the experiment was to check if the noises are stimulated by the sounder transmitter [Danilkin and Pulinets, 1989]. The transmitter was switched off and the two consecutive ionograms with intervals of 8 s are presented in Figure 15. It is evident that after switching off the transmitter the noises did not disappear, so they indicate the turbulent region of the high-latitude ionosphere where they are generated naturally, primarily due to particle precipitation. So the monitoring of wide-band noises in the DNT resonance frequency band gives a possibility to make a diagnosis of the state of the ionosphere. This technique is in the very beginning but the hope exists that it will be very useful.

The natural noises may be useful, not only in the DNT frequency band but in the wide range of frequencies. For this purpose the voltage of the AGC is used which is measured before the ionosonde transmitter pulse. Following the registered intensity level along the satellite orbit, many ionospheric properties can be observed. One can see the different ionospheric structures fitted by the upper hybrid resonance cutoff (Figure 16). The increasing of intensity in the whistler frequency range, as well as in the upper-hybrid range, in the polar region confirm electron precipitation as the source of HF emission in that area. The envelope of signals from ground-based broadcasting transmitters indicate the changes in f_oF2 frequency. The form and position of the main trough within the F-layer maximum could be determined. In some sense this representation is more rich than the N(h) profiles because it gives information not only on the electron density, but on a lot of physical processes taking place in the ionosphere. So measurement of natural noises is an additional means of topside sounding [Pulinets et al., 1988].

4.3 Galactic noise as a source of local density estimation

In the frequency range between the local upper hybrid frequency f_{UHR} and critical frequency f_oF2, the principal source of natural radio noise is galactic. If the sensitivity of the sounder receiver is sufficient enough to register them, it is possible to use the low frequency cutoff

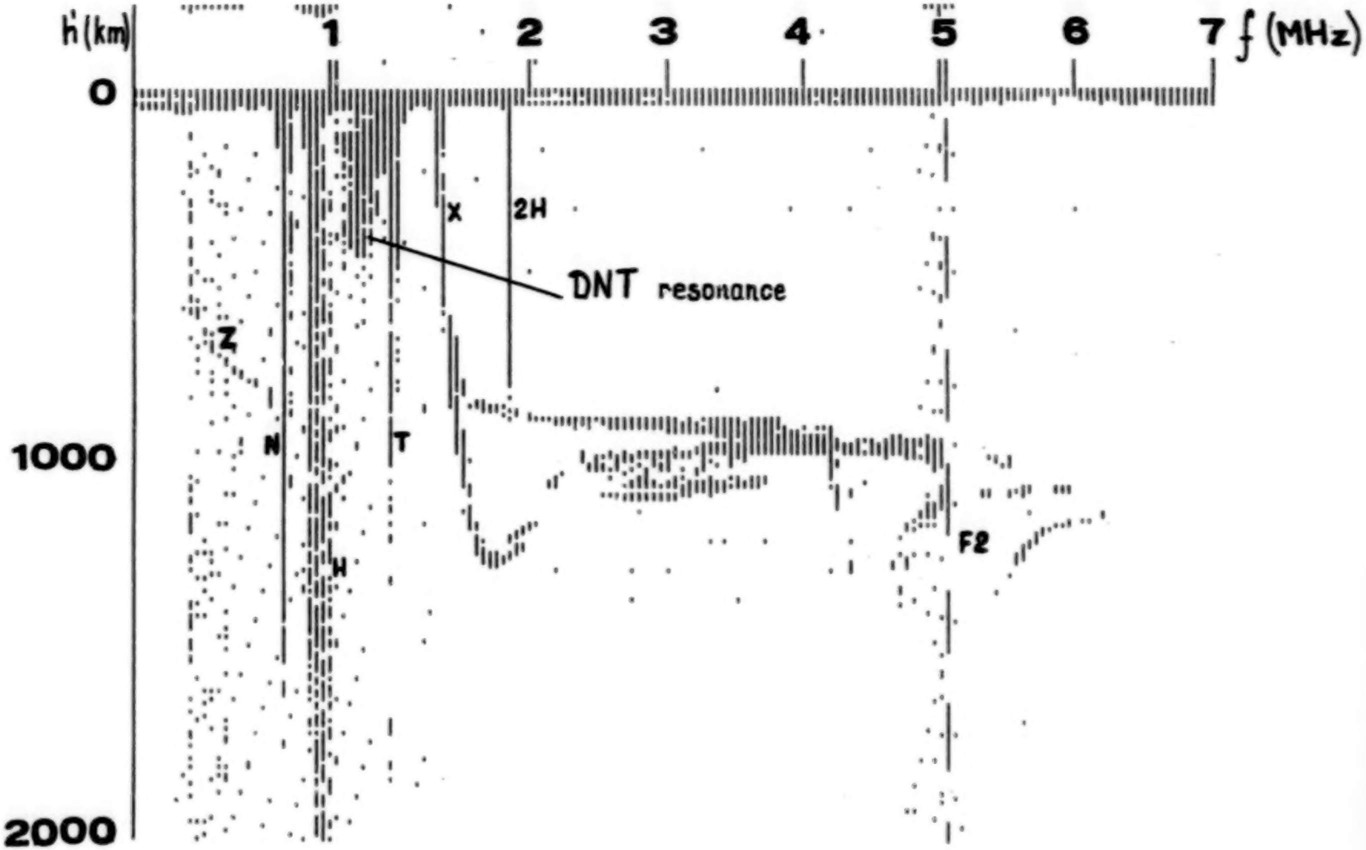


Figure 13. DNT resonance on the topside ionogram.

117

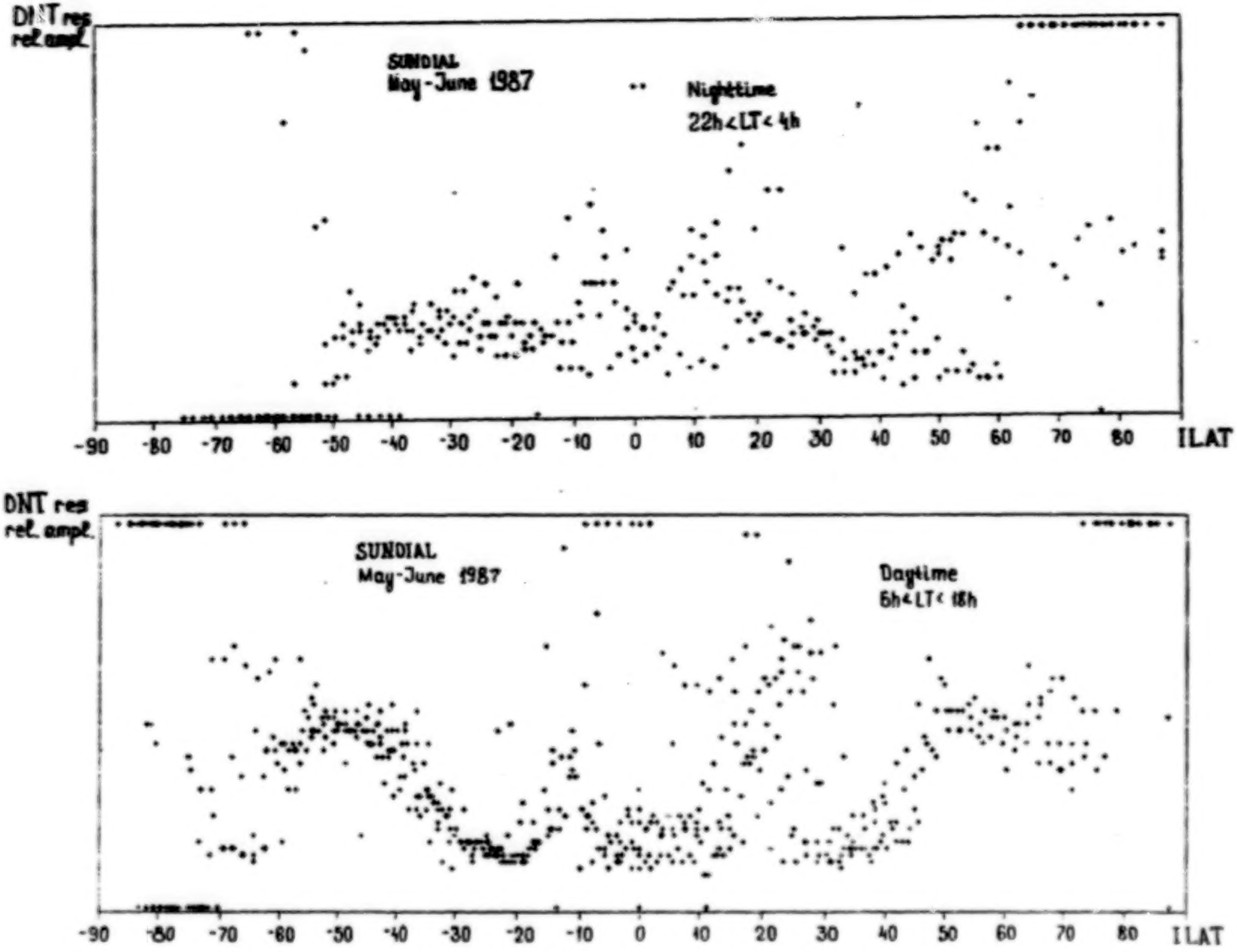


Figure 14. Global distribution of DNT resonance amplitude.

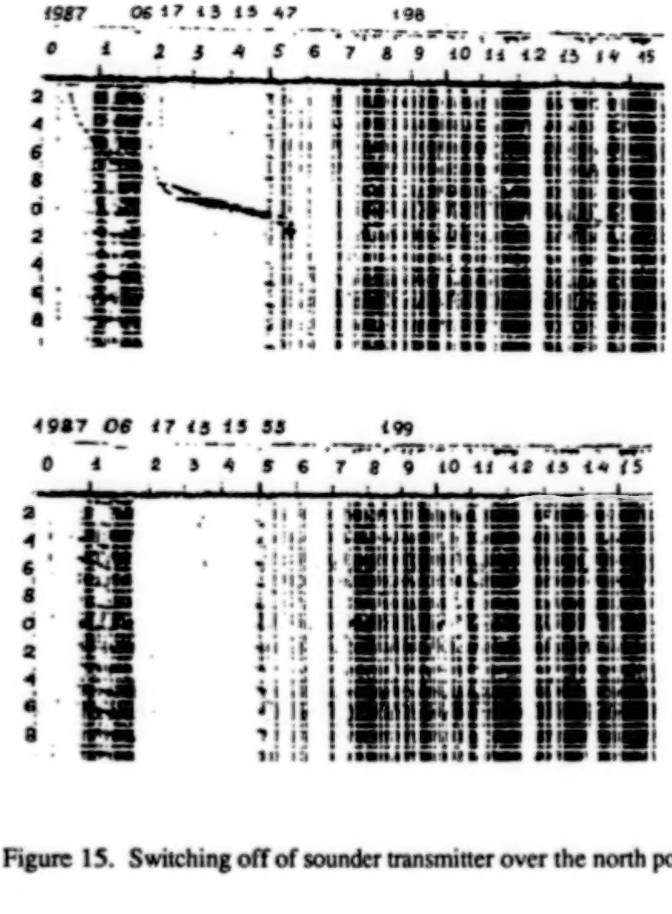
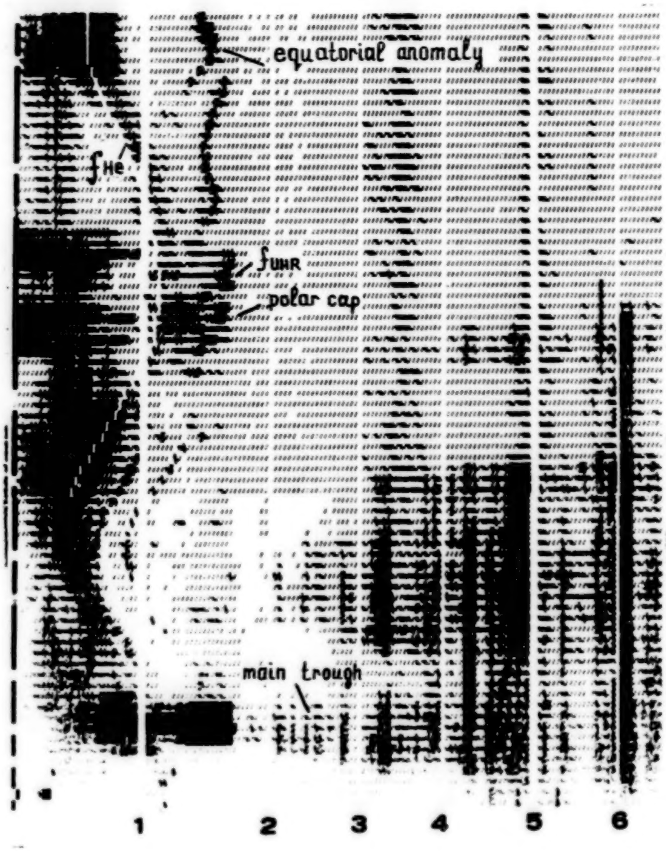


Figure 15. Switching off of sounder transmitter over the north polar cap.

COSMOS-1809 R: 735 DATE: 09-02-1987 UT 09:47:33

H: 960-995 km



LAT	LONG	L	MLT
13.42	83.76	1.08	15.41
-19.72	85.80	1.50	15.50
-56.02	91.59	8.34	15.55
-81.90	191.39	18.93	03.12
-49.70	246.25	2.02	03.39
-13.37	250.63	1.17	03.42
23.33	252.95	1.58	03.41
59.63	259.90	9.38	03.32

Figure 16. AGC dynamic spectra from COSMOS-1809 data.

of the galactic noise to estimate the local upper hybrid frequency, which was suggested by designers of the ISS-B satellite [Summary plots..., 1983]. In the ionosphere above the F-layer peak, there is a cone of directions within which an extraterrestrial noise can reach the satellite. The cone is narrower on the lower frequencies and the AGU level which integrates over the whole cone becomes lower on lower frequencies. The scaling of the cutoff frequency f_{CNO} gives estimation value for local plasma frequency. It is shown in Figure 17 a good correlation between the frequencies which is better than correlation between f_{CNO} and f_{UHR} and between f_{CNO} and f_x, nevertheless that f_{CNO} is always higher than f_{UHR} and distributed near the f_x cutoff. Naturally this technique could be used when the different kinds of interference are absent and the local density is higher or of order of local gyrofrequency.

4.4 Transionospheric sounding

The topside sounder data do not contain information on the internal ionosphere which determines the conditions for RF signal propagation. The technique of transionospheric sounding gives a possibility to accomplish the topside sounding information by using the frequency band close to the critical frequency f_0F2. The method was proposed first by Danilkin [1985, 1987b]. The main idea of the technique is to record the time delays of sounder pulses on frequencies, not only in the frequency band lower than the critical frequency f_0F2, but on the higher frequencies also. The higher frequency signals go through the bulk of the ionosphere and are named "transionospheric" signals. It is necessary to have ground-based equipment synchronized with the topside sounder sweep. The synchronizing signals may be transmitted by the additional line not so affected by ionospheric layers, for example on frequencies higher than 100 MHz. Two almost equivalent methods are used [Danilkin, 1987b]: registration of the sounder transmitter signals on frequencies higher than f_0F2 by synchronized ground-based transmitter signals by the topside sounder receiver (Figure 18b) -- reverse transionospheric sounding (RTIS). These methods give cross sections of the ionosphere within 5000 km radius. The total oblique profiles could be calculated. The problem comes in determining the function of the electron concentration from the equation

$$p' = \int_E^{sat} \mu'(\psi)ds = \int_E^{sat} \mu'(\psi) \cos\alpha(\psi)ds$$

where sat and E are the coordinates of the satellite and ground-based station, respectively:

$$\psi = [t, f_H\, \theta(S), \varphi_N(S)]$$

μ' is the group refraction index, f_H the electron gyrofrequency, f the wave frequency, θ angle between the local vertical and geomagnetic field, α angle between the wave and ray normals.

The method gives a possibility to determine accurately the maximum used frequencies (MUF), to estimate the horizontal gradients of electron concentration. One of the possible applications of the method is the diagnosis of ionospheric irregularities schematically shown in Figure 19. The bays shown on transionograms 1 and 2 between frequencies f_1 and f_2 manifest the presence of irregularities between ray traces of the waves with these frequencies. But if we take only one ionogram for position of satellite 1 the ambiguity exists on the position of the irregularity (the upper or lower ionosphere). But comparing with position number 2 decides the problem. The difference between frequencies f_1 and f_2 determines the irregularity dimension and the $\Delta D'$ shown on transionograms determines the deviation of concentration $\pm \Delta N$ in the irregularity.

5. Conclusion

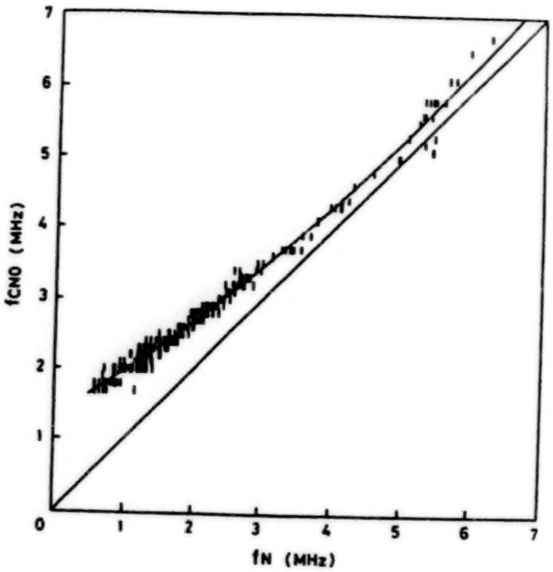


Figure 17. Empirical relationship between f_n and f_{CNO}.

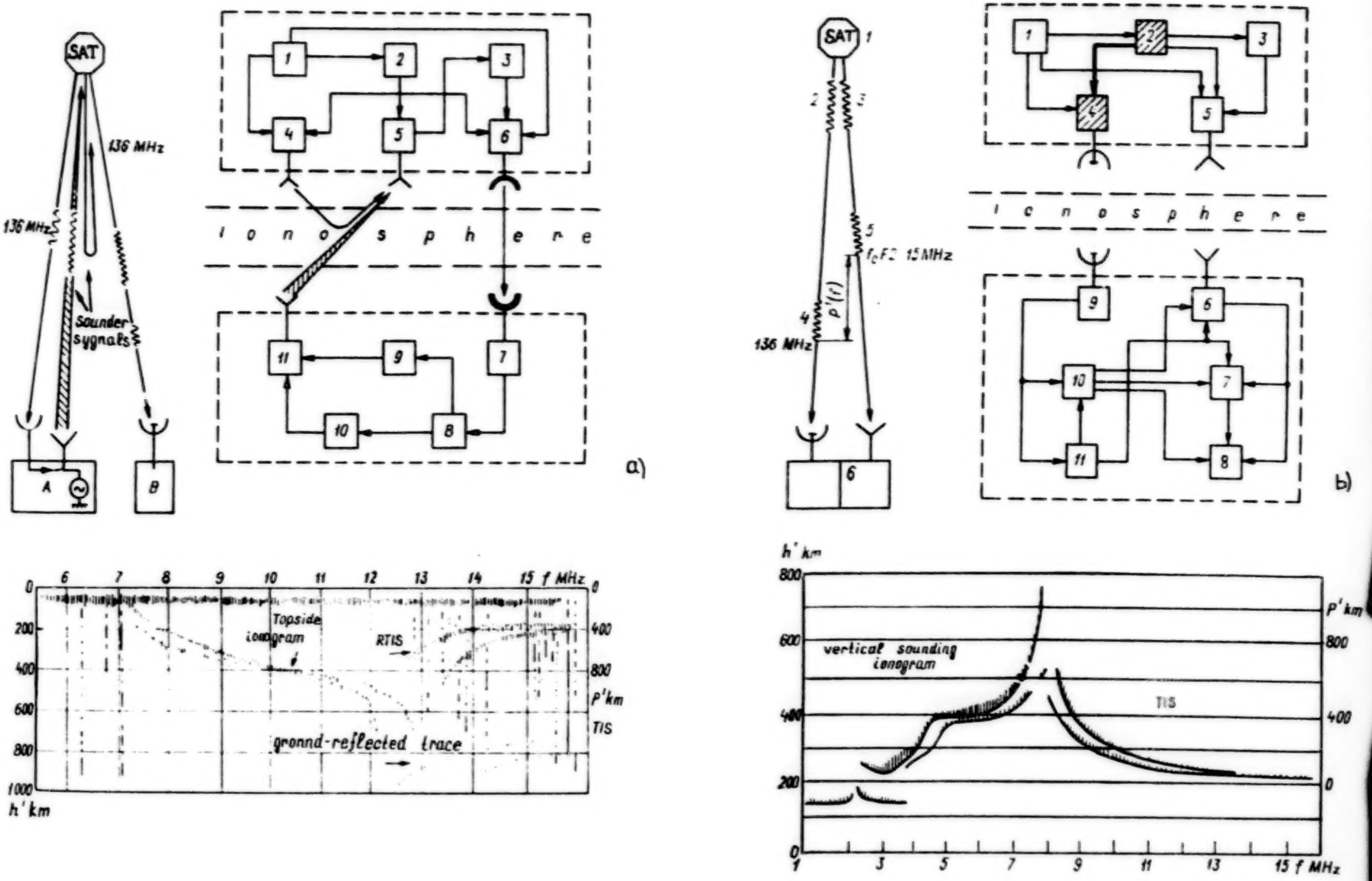


Figure 18. Chart diagrams of a) transionospheric (TIS) and b) reverse transionospheric (RTIS) sounding.

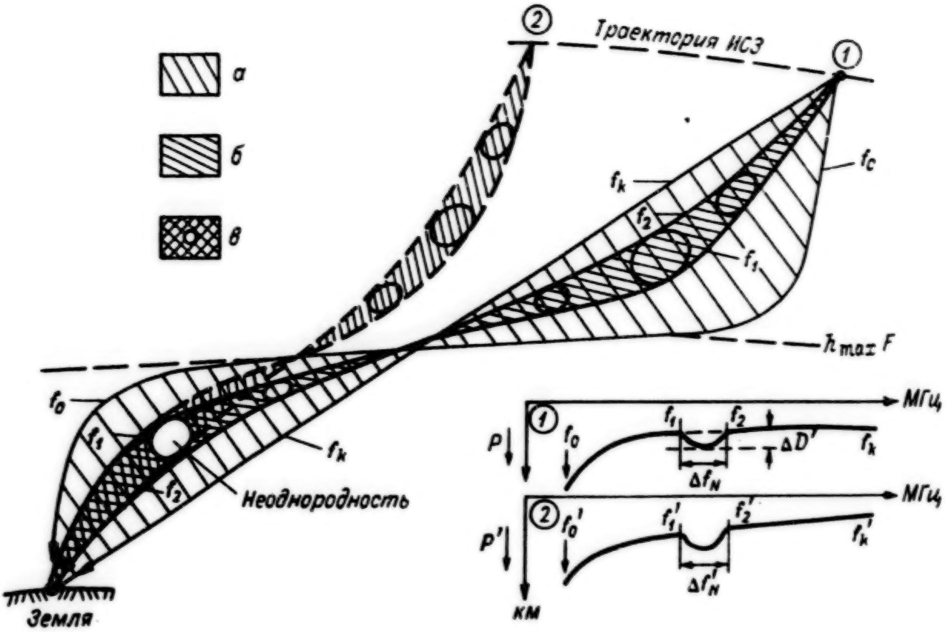


Figure 19. Diagnostics of ionospheric irregularities by transionospheric sounding.

Only the simple mention of the problems and new techniques in topside sounding took a lot of space. What are the key problems for complete knowledge of the ionosphere with the help of the method? I'll try to summarize them.

1. The development of N(h) profile calculation methods with correction taking into account the traces of signals reflected from the Earth and process of of F-layer formation.
2. Development of the onboard processing techniques.
3. Creation of a multisatellite system of global monitoring of the ionosphere.
4. Accomplishment of the topside sounding by additional closely related techniques such as use of information on the natural noises and plasma resonances, transionospheric sounding.
5. More wide use of topside sounder data for ionospheric modeling and forecast and creating more sophisticated global ionospheric models.

6. References

Austen, J. R., S. J. Franke, and C. H. Liu, Ionospheric imaging using computerized tomography, *Radio Sci., 23*, 299, 1988.

Atlas of Ionospheric Critical Frequency (f_oF2) obtained from Ionospheric Sounding Satellite-b Observation, issued by Radio Research Laboratory Ministry of Posts and Telecommunications, Japan, 1979.

Benson, R. F., Ionospheric plasma resonances: time durations vs latitude, altitude and F_Nf_H, *Planet Space Sci., 20,* 683, 1972.

Benson, R. F., Stimulated plasma instability and nonlinear phenomena in the ionosphere, Radio Sci., 17, 1637, 1982.

Benson, R. F., Auroral kilometric radiation: Wave modes, harmonics, and source region electron density structures, *J. Geophys. Res., 90,* 2753, 1985.

Benson, R. F., and W. Calvert, ISIS 1 observation in the source of auroral kilometric radiation, *Geophys. Res. Lett., 6*, 479, 1979.

Ben'kova, N. P., M. G. Diominov, A. T.Karpachiov, N. A. Kochenova, Yu. V. Kushnerevsky, V. V.Migulin, S. A. Pulinets, and M. D. Fligel, Longitude features shown by topside sounder data and their importance in ionospheric mapping, *Adv. Space Res., Ionospheric Informatics*, 1988, in press

Bratsun, D. S., P. F. Denisenko, N. A.Zabotin, S. A. Pulinets, and V. V. Selegey, Peculiarities of the topside ionograms in presence of small-scale irregularities in the ionosphere, Preprint IZMIRAN 44(798), to be published in *Geomagn. Aeron.*, 1988.

Danilkin, N. P., Radioscopy of the ionosphere on the transparency range border, *Radiotekhnika, 9*, 3, 1985.

Danilkin, N. P., Systematical sounding -- the basis for ionospheric service building, in *Ionospheric-Magnetospheric Service*, Gidrometeoizdat, Leningrad, 46, 1987a.

Danilkin, N. P., Transionospheric radio sounding as a means for ionospheric conditions control, in *Ionospheric-Magnetospheric Service*, Gidrometeoizdat, Leningrad, 9, 1987b.

Danilkin, N. P., and S. A. Pulinets, On the nature of wind-band noises by the topside sounding data in high latitudes, in *Results of the First Scientific Expedition on Atomic Ice-Breaker Sibir in Near-Pole Region, AANII*, Gidrometeoizdat, Leningrad.

Denisenko, P. F., N. A. Zabotin, S. A. Pulinets, and V. V. Selegey, Transformation of ordinary waves into extraordinary waves according to data from topside sounding of the ionosphere, *Geomagn. Aeron., 27*, 473, 1987.

Diominov, M. G., and A. T. Karpachiov, The longitudinal effect in main trough configuration 1 and 2, *Geomagn. Aeron., 26*, 63 and 682, 1986.

Diominov, M. G., A. T. Karpachiov, B. B. Afonin, I. P.Khar'kov, and J. Schmilauer, The longitudinal control of the electron temperature in the subauroral ionosphere, *Geomagn. Aeron., 27*, 409, 1987.

Fligel, M. D., On the one possible mechanism of oblique propagation of radio waves in topside ionosphere, *Geomagn. Aeron., 29*, 71, 1989.

Franklin, C. A., and M. A. Maclean, The design of swept-frequency topside sounders, *Proc. IEEE, 57*, 897, 1969.

Gulyaeva, T. L., Global reconstruction of the ionospheric F-region structure, *Ind. J. Radio Space Phys., 17*, 155, 1988.

Gurnett, D. A., F. L. Scarf, R. F. Fredricks, and E. J. Smith, The ISEE-1 and 2 plasma wave investigation, *IEEE Geosci. Electr., GE-16*, 225, 1978.

Hagg, E. L. E. J. Hewens,and G. L. Nelms, The interpretation of topside sounder ionograms, *Proc. IEEE, 57*, 949, 1969.

Huang, X., and B. W. Reinisch, Automatic calculation of electron density profiles from digital ionograms. 2. True height inversion of topside ionograms with the profile-fitting method, *Radio Sci., 17*, 837, 1982.

Igi, S., and K. Aikyo, Automatic identification of resonance spikes for ionograms obtained by ionospheric sounding satellite-b (ISS-b), *J. Radio Res. Lab., 33*, 169, 1986.

Jackson, J. E., The reduction of topside ionograms to electron-density profiles, *Proc. IEEE, 57*, 976, 1969.

James, H. G., Direction-of-arrival measurements of auroral kilometric radiation and associated ELF data from ISIS 1, *J. Geophys. Res., 85*, 3367, 1980.

Karpachiov, A. T., The global longitudinal effect in the nighttime upper ionosphere by the Intercosmos-19 satellite data, Ph.D. Thesis, IZMIRAN, 1987.

Kochenova, N. A., and M. D. Fligel, On the using of ground-reflected trace on the topside ionograms for more precise N(h)-profile calculation, *Ionospheric Researches*, in press, 1989.

Kovaliov, V. A., V. V. Sotsky, and M. D. Fligel, Improving of accuracy of determining the N(h) profile by the topside sounding data, paper presented at National conf. on Ionogram Interpretation, Troitsk, June 1989.

Mathwich, H. R., D. E. Aubert, A. F. Martz, K. Bibl, B. W. Reinisch, and D. Lewis, An advanced mission to map the world-wide topside ionosphere, paper presented at Symp. on the Effects of the Ionosphere on Radio Systems, Naval Res. Lab., Office of Naval. Res., and Air Force Geophys. Lab., Alexandria, VA, 1981.

Muldrew, D. B., Nonvertical propagation and delayed-echo generation observed by the topside sounders, *Proc. IEEE, 57, 1097*, 1969.

Nishizaki, R., N.Matuura, and A. Takahira, Topside sounder onboard ISS-b, *Rev. Radio Res. Labs., 28,* 155, 1982.

Oya, H., Sequence of diffuse plasma resonances observed on Alouette 2 ionograms, *J. Geophys. Res., 75*, 4279, 1970.

Oya, H., Verification of theory on weak turbulence relating to the sequence of diffuse plasma resonances in space, *Phys. Fluids, 14*, 2487, 1971.

Oya, H., A.Morioka, and T.Obara, Leaked AKR and terrestrial hectometric radiations discovered by the plasma wave and planetary plasma sounder experiments onboard the Ohzora (EXOS-C) satellite - Instrumentation and observation results of plasma wave phenomena, *J. Geomagn. Geoelectr., 37,* 237, 1985.

Pulinets, S. A., P. F. Denisenko, N. A. Zabotin, and T. A. Klimanova, New method for small-scale irregularities diagnostics from topside sounder data, paper presented at SUNDIAL Workshop, McLean, VA, May 1989.

Pulinets, S. A., A. Kiraga, and Z. Klos, Natural noises and ionospheric mapping, Adv. Space Res., *Ionospheric Informatics*, in press, 1988.

Pulinets, S. A., and V. V. Selegey, Ionospheric plasma modification in near satellite reg ion during topside sounding by powerful radio pulses, Preprint IZMIRAN, 4,(415), 25, 1983

Pulinets, S. A., and V. V. Selegey, Ionospheric plasma modification in the vicinity of a spacecraft by powerful radio pulses in topside sounding, *J. Atmos. Terr. Phys., 48*, 149, 1986.

Reinisch, B. W., and X. Huang, Automatic calculation of electron density profiles from digital ionograms, 1. Automatic O and X trace identification for topside ionograms, *Radio Sci., 17*, 421, 1982.

Serebryakova, M. V., The topside ionograms processing, Preprint IZMIRAN, 62,(595), 20, 1986.

Sotsky, V. V., The height variations of electron concentration in the ionosphere on the ground-based, topside and transionospheric sounding data, Ph.D. thesis, 1986.

Summary Plots of Ionospheric Parameters Obtained from Ionosphere Sounding Satellite-B, issued by Radio Research Laboratories Ministry of Posts and Telecommunications, Japan, 1 - 4, 1983.

Szuszczewicz, E. P., B. Fejer, R. Schunk, E. Roelof, and R. Wolf, SUNDIAL: An international program in solar-terrestrial physics, *WITS Handbook, 1,* 208, 1988, eds. C. H. Liu and Belva Edwards, SCOSTEP Secretariat, University of Illinois at Urbana-Champaign.

Topside Sounding and the Ionosphere, Special Issue *Proc. IEEE, 57,* 1969.

Vasiliyev, G. V., L. P. Goncharov, Yu. V. Kushnerevsky, V. V. Migulin, and M. D. Fligel, The satellite system IS-338 for pulsed sounding of the ionosphere, in *The Apparatus for Investigation of the Topside Ionosphere,* Moscow, 13, 1980a.

Vasiliyev, G. A., A. A. Kalichev, Yu. A. Kapitonov, V. N. Knyazev, Yu. M. Kovalyov, E. V. Pogoda, V. P(. Polyanski, and V. L. Rosin, The satellite ionospheric station ION-1, in *The Apparatus for Investigation of the Topside Ionosphere,* Moscow, 30, 1980b.

Vasiliyev, G. V. V. I. Dmitriev, L. A. Emeliyanova, Yu. V. Kushnerevsky, M. D. Kulagin, V. V. Migulin, M. D. Fligel, I. P. Khar'kov, Yu. N. Shaulin, L. D. Shoya, and A. F. Yakovlev, On the same class of problems in the solar-terrestrial coupling investigation, *Doklady of USSR Acad. Sciences, 257,* 1117, 1981.

Vidmar, R. J., and F. W. Crawford, Long-delayed radio echoes: mechanisms and observations, *J. Geophys. Res., 90,* 1523, 1985.

Warnock, J. M., Sideband structure observed by topside sounders, *Proc. IEEE, 57,* 1135, 1969.

Chapter 4

SCINTILLATION TECHNIQUE FOR PROBING IONOSPHERIC IRREGULARITIES

Santimay Basu
Geophysics Laboratory (LIS)
Hanscom AFB, MA 01731

Sunanda Basu
Physics Research Division
Emmanuel College
Boston, MA 02115

ABSTRACT

The phenomenon of radio wave scintillation caused by ionospheric irregularities of electron density are described. The different indices of scintillation that provide a measure of intensity and phase scintillation have been defined. Both simple and sophisticated receiving systems used for recording scintillations are briefly described. The use of scintillation experiments in exploring the physical mechanisms of irregularity formation and in defining the constraints of practical communication systems are outlined.

INTRODUCTION

Radio waves from satellites or radio stars during their passage through the ionospheric irregularities of electron density develop random phase fluctuations across the wavefront. As the wavefront travels towards the ground, phase mixing occurs and, as a result, not only phase but amplitude fluctuations as well, develop on the ground. Due to the relative motion between the satellite, the ionosphere and the receiver on ground, the spatial pattern of amplitude and phase variation sweeps past the receiver and temporal variations of phase and amplitude known as scintillations are recorded by the receiver. In the case of radio stars or geostationary satellites, the temporal variation is caused by the ionospheric drift whereas in the case of low altitude (≈1000 km) orbiting satellites, the satellite motion projected on to the ionosphere dictates the temporal structure.

SCINTILLATION PARAMETERS

The amplitude and phase fluctuations of the recorded signal are statistically characterized by two major parameters, amplitude and phase scintillation indices, denoted respectively by S_4 and σ_ϕ. Additional parameters, such as the signal decorrelation time and spectral shape of signal fluctuations are also of importance for specifying effects on systems.

Scintillation Index, S_4

The amplitude scintillation index, S_4, is defined as the ratio of the standard deviation of signal intensity and the average signal intensity and defined as

$$S_4 = \frac{(<I^2> - <I>^2)}{<I>^2} \tag{1}$$

The time interval over which this parameter is computed depends on the time period over which the signal fluctuations are observed to be stationary. For geostationary satellite observations, a 3-minute to 15-minute data segment has been considered to be optimum whereas for orbiting satellites a data segment varying between 10-30 seconds has been used.

RMS Phase Deviation, σ_ϕ

The phase scintillation index is defined as the standard deviation of a linearly detrended phase data segment. The linear detrending over the appropriate data segment discussed in the previous paragraph removes the background phase variation caused by smooth changes of ionization density. Thus the phase scintillation index is defined as

$$\sigma_\phi = (< \phi^2 > - < \phi >^2)^{1/2} \tag{2}$$

Intensity Decorrelation Time, τ_I

In addition to the above two parameters the intensity decorrelation time, τ_I, defined as the time shift required to obtain signal correlation of 0.5 is also important. The decorrelation time, τ_I, is defined as

$$\frac{< I(t)\, I(t + \tau_I) >}{< I(t) >^2} = 0.5 \tag{3}$$

In the case of receiving systems capable of acquiring both signal amplitude and phase from a satellite source, the complex decorrelation time can be obtained by replacing the intensity terms in Eq. (3) by the complex amplitude.

Scintillation Spectrum

The time series of intensity and phase scintillation can be analyzed by the standard Fast Fourier Transform (FFT) or Maximum Entropy Method (MEM) to derive a plot of the variation of the power spectral density of amplitude or phase as a function of fluctuation frequency. The length of the data sample chosen for such spectral analysis needs to be stationary. For geostationary satellite observations a 3-minute scintillation data sample is, in general, found to be optimum. In such applications, the data digitization rate is required to be in the range of 30-50 Hz. The frequency range of the spectral analysis is dictated at the low frequency end by the inverse of the data sample duration and at the high frequency end by the Nyquist frequency which is half of the digitization rate. Thus for a 180-sec data segment from a geostationary satellite and 50-Hz sampling frequency, the frequency interval over which the spectrum is obtained corresponds to .0056 Hz to 25 Hz. For an average ionospheric drift of 50 ms^{-1}, the above frequency interval corresponds to spatial irregularity wavelength range of 8.9 km to 2 m. In the equatorial region, when strong to moderate levels of scintillations are observed, the ionospheric drift typically varies smoothly between 200 ms^{-1} to 50 ms^{-1} in the course of an evening which changes the coverage of irregularity wavelengths by a factor of 4. At high latitudes, the ionospheric drift variations are more severe with drifts changing from 50 ms^{-1} to 2000 ms^{-1} in a matter of minutes. Thus the irregularity wavelengths covered by the spectral analysis can vary greatly at high latitudes.

For orbiting satellites, at an altitude of 1000 km, the scan velocity of the propagation path through the F-region height of 300 km corresponds to 2.2 km/sec. The temporal structure of scintillation is, therefore, dictated by the scan velocity except at high latitudes where the ionospheric drift may not, on occasions, be negligible. In the case of orbiting satellites, data length of 10 sec and digitization rate of 125 Hz is found to be optimum. Considering only the ionospheric scan velocity, the above parameters provide an irregularity wavelength coverage of 22 km to 18 m.

The scintillation parameters, defined above, can be readily computed from digital recordings. Many scintillation recording systems still employ chart recorders instead of digital recorders. For such systems, Whitney et al. [1969] defined the amplitude scintillation magnitude, SI_{dB}, in terms of dB excursions between the third peak up from the minimum and third peak down from the maximum signal levels. By providing calibrations with a signal source applied to the antenna input terminals of the receiver in terms of dB and using relatively fast chart speed so that the individual fadings and signal enhancements can be delineated, the third peak method is a useful way to quantify the analog recordings. Comparing digital and analog recordings Whitney et al. [1969] showed that the third peak method provides the dB excursions between the 2nd and 98th percentiles of signal fluctuations. In addition, he provided a useful graph that can be used to convert SI (dB) indices to S_4 indices. Attempts have also been made to manually digitize fast chart recordings to derive scintillation spectra. Reliability of such spectra is, however, limited because the response time of chart recorders does not exceed a few Hz thereby limiting the Nyquist frequency to 1 Hz or less. Since strong scintillations often contain fluctuation frequencies that exceed 10 Hz, manually derived spectra often provides misleading results.

SCINTILLATION RECORDING SYSTEMS

Total Power Recording System

The total power or systems recording intensity scintillations from geostationary satellites transmitting circularly polarized signals in the VHF range (typically at 136, 137, 244 and 257 MHz) can be achieved by a simple system. It consists of a ten element Yagi antenna, a commercial converter (noise figure of 3 dB, conversion gain of 10 dB) that converts VHF frequencies to a radio frequency (typically 28 MHz or 10.7 MHz) followed by a communications receiver tuned to accept the converted frequency, and operated with an IF bandwidth of 4 or 8 KHz. The above IF bandwidth is found to be wide enough to be unaffected by the frequency drift of the local oscillator. The communication receiver is operated in the fast AGC mode providing a time constant of .01 sec. The AGC voltage is fed to a chart recorder. In order to protect the pen from jitter, an external RC circuit with a time constant of 0.1 to .05 sec is often inserted between the AGC port and the pen recorder. The pen recorder is usually equipped with a d.c. buckout facility. The signals from geostationary satellites received with a ten-element Yagi antenna usually develop 110 dBm at the converter input. In order to calibrate the system, the antenna is disconnected from the converter and a calibrating signal source (providing 100-120 dBm at VHF followed by attenuators capable of varying the attenuation over a 100-dB range in 1-dB steps) is connected to the converter input. This is used to obtain calibration levels in the chart recorder at 1 to 3-dB intervals over the full scale deflection range. The system is so adjusted that the satellite signal level runs at a level 6 dB below the full-scale deflection. It should be noted that for scintillations arising from a diffraction process, the positive excursions do not exceed a 6-dB level although the associated negative excursions may approach 25 dB.

Recording of amplitude scintillations of L-band signals from stationary satellites can be achieved with a similar system by replacing the Yagi antenna with an 8-ft paraboloidal antenna and a low noise L-band converter.

With the advent of desktop computers, the AGC signal after d.c. buckout can be digitally recorded on a disk and processed to obtain the full range of statistical parameters, such as, S_4 index, decorrelation time , and scintillation spectrum, on a real-time basis.

Phase and Amplitude Scintillation Recording from Stationary Satellites

Stationary satellites transmitting single frequency phase coherent signals have been used to record both amplitude and phase scintillations by using a sophisticated computer-controlled receiver.

For such measurements an extended dynamic range receiver with an extremely stable local oscillator is employed. The receiver operates under computer control and once tuned to within a few hertz of a signal detected in a 10-Hz bandwidth, self-tunes to within ±1 millihertz of the mean frequency as determined by the zero crossings averaged over a 20-sec period. Subsequent changes in frequency, either due to changes in ionospheric or geometrical doppler, are sensed by the system, which then retunes. At each retune, the local oscillator frequency information is recorded to allow reconstruction of the long-term phase in subsequent processing. In this way the system can measure signal phase variations with precision as would a coherent system except for any long-term relative frequency drifts between the satellite and receiver references.

Once a signal is properly acquired by the receiver, its quadrature components are sampled at 10 Hz and are digitally recorded along with time and pertinent system information. During initial off-line processing, these data are converted to signal intensity and continuous phase. While simple in concept the generation of continuous phase over long observation periods is prone to numerical difficulties. Since this is a single frequency measurement, the accumulation of phase over a period of several hours can result in extremely large values. These problems have been avoided by calculation and removal of the largest scale (for example, few hours) dispersive doppler changes during pre-processing.

Following initial processing, only phase variations with periods shorter than some tens of minutes remain. The data can then be treated using methods similar to those designed for, and proven, during the Wideband experiment [Fremouw et al., 1978]. Basically, this consists of separation of rapidly varying scintillation components of the signal from the longer term trends. The spectral components are separated by passing the phase data through a sharp cutoff high pass digital filter ($f_c = 0.0067$ Hz); there is generally no need to filter signal intensity, which has no low frequency component in the constant signal level, geostationary case.

PHASE AND AMPLITUDE SCINTILLATION MEASUREMENTS WITH MULTIFREQUENCY BEACON TRANSMISSIONS

DNA Wideband Satellite

Complex signal scintillation measurements have been performed by using multifrequency coherent radio beacon transmissions from an orbiting satellite. In 1976, the U.S. Defense Nuclear Agency launched the DNA Wideband satellite in a 1000-km orbit which radiated ten coherent radio spectral lines between the VHF and S bands [Fremouw et al., 1978]. The ten spectral lines were derived from a fundamental frequency of 11.4729 MHz and corresponded to one VHF, seven UHF, one L-band and one S-band transmission. The receiving system consisted of a tracking antenna and a receiver which was designed to maintain the coherence of the transmitted signals. The receiver incorporated a frequency synthesizer which generated a replica of the transmitted signal spectrum which were offset by the intermediate frequencies to generate the local oscillator signals.

The synthesizer was phase-locked by a loop operating on the output of the S-band reference receiver. Except for the first mixer, the reference receiver was identical to the measurement receiver channels at various frequencies. Nine measurement receiver channels were employed; one for VHF , one each for seven UHF channels, and one for L band. Remote receiving antennas comprised of identical measurement channels were used to perform spaced receiver scintillation observations.

The signals in each receiver channel were finally translated to an essentially zero frequency baseband by means of two quadrature detectors. The receiving system thus provided coherently detected in-phase and quadrature components at VHF, seven UHF, and one L-band channel. Thus phase and intensity scintillation indices at all these frequencies, as well as, the second order statistics of complex signal scintillation in the temporal, spatial and spectral domains could be obtained. In addition to scintillation measurements, the second difference of phase measurement at the three UHF comb frequencies provided the total electron content (TEC) of the ionosphere.

The second difference of phase, $\Delta_2\phi$, is obtained by first forming the phase difference between a carrier and its upper sideband (413 MHz and 435 MHz) and the same carrier and its lower sideband (390 MHz and 413 MHz) and then taking the difference of the two differences. It can be shown [Fremouw et al., 1978] that

$$\Delta_2\phi = \frac{e^2}{2\pi\varepsilon_o m c}\frac{f_m^2}{f^3}N_T$$

where, e - electronic charge,
m - mass of an electron,
ε_o - permittivity of free space,
c - speed of light,
f_m - the frequency separation between the UHF triad,
f - center UHF frequency of the triad,
N_T - total electron content up to satellite height.

The full complement of Wideband satellite signals were monitored from the auroral location of Poker Flat, Alaska, and the equatorial locations of Ancon, Peru and Kwajalein, Marshall Island. A brief period of Wideband satellite measurement was initially performed at Stanford, California through the relatively undisturbed ionosphere. The satellite was placed in a sun-synchronous orbit so that equatorward crossings were obtained around 1120 and 2320 local time and the auroral ionosphere was intercepted in the post-midnight time frame. The nighttime equator crossings could have been a couple of hours earlier for the interception of most intense early evening irregularities.

DNA HiLat Satellite

The DNA HiLat satellite was launched on June 27, 1983 in a near circular 800-km orbit with an inclination of 82°. The satellite carried a beacon transmitter package similar to the Wideband satellite, but, in addition, performed satellite in-situ measurements which were telemetered to ground using the L-band transmission of the beacon. The objective of the mission was to characterize the physical environment of the upper atmosphere through in-situ measurements, establish their relationship with the generation of plasma density irregularities and measure the effects of the irregularities on radio wave propagation. The beacon transmitter provided coherent transmission at L-band (1239 MHz), three UHF frequencies (447, 413 and 378 MHz), and one at VHF (138 MHz). The L-band signal while serving as a phase reference for phase scintillation measurement was also modulated by the telemetry signal. At the receiver, an L-band carrier recovery loop has been used to generate the reference signal for the demodulation of the telemetry data and synchronous demodulation of the UHF and VHF frequencies.

The HiLat satellite in-situ instrument package consisted of a J-sensor for the measurement of the local flux of electrons over the energy range of 20 eV to 20 keV in 16 channels, a magnetometer for the measurement of field-aligned 'Birkeland' currents, an ion drift meter for the measurement of thermal ion density and cross track ion drift, a retarding potential analyzer for the measurement of the ram component of ion drift, ion composition and ion temperature and an electron Langmuir probe. In addition, a vacuum ultraviolet (UV) imager was included to obtain auroral UV images in full daylight in selectable spectral windows over the range of 1100 Å to 2078 Å. The in-situ package provided an excellent array of instruments that characterized the currents, particles and fields in the high latitude ionosphere.

The data from the HiLat satellite were acquired at Sondrestrom, Greenland located in the dayside cusp region and at Fort Churchill, Canada and Tromso, Norway corresponding to two auroral locations in the American and European sectors. In addition, a portable Rover receiving system normally located at Bellevue, Washington could be deployed anywhere at high latitudes to perform campaign observations.

The Polar Bear Satellite

This satellite carrying a beacon package identical to that of the HiLat satellite but more limited in-situ instrumentation providing UV imaging and magnetometer measurements only was launched in 1986. Both HiLat and Polar Bear satellites were monitored with identical receiving systems at the four high latitude locations mentioned above.

APPLICATION OF SCINTILLATION MEASUREMENTS

Scintillation measurements have been used as a key diagnostic for the presence of the irregularities of electron density in the ionosphere. This knowledge is important, on one hand, to determine the spatial and temporal distribution of the ionospheric irregularities and to understand the physical processes that lead to the formation of such irregularities. On the other hand, the characteristics of the scintillating signal structure determines how the performance of a communication link or a radar system can be degraded. The key scintillation characteristics that contribute to our knowledge of geophysics and the parameters that dictate the degradation of system performance are briefly summarized in the following paragraphs.

The key irregularity parameter, namely the integrated electron density deviation (ΔN) of the ionospheric irregularities, can be derived from intensity and phase scintillation measurements. In the framework of weak scatter theory the intensity and phase scintillation indices, S_4 and σ_ϕ, have been related [Rino, 1979] to ΔN and the geometrical term accounting for both the geometry of the propagation path and the irregularity anisotropy. Of the two scintillation indices, the phase scintillation index, σ_ϕ, is more useful as it is related, in a straightforward manner, to the irregularity parameters.

The other irregularity parameter of great importance to our understanding of the generation of irregularities is related to the form and shape of the irregularity spectrum. The scintillation spectrum reflects the form of the irregularity spectrum and the spectral index of scintillation has been simply related to the 3-dimensional irregularity spectral index. As such, the spectral studies of both intensity and phase scintillations have been performed and important conclusions regarding the irregularity spectral shapes, spectral indices and drifts have been derived [Basu et al., 1985].

Scintillation measurements, because of their relative simplicity, have been most useful in determining global morphology of irregularities. Earlier reviews were presented by Aarons [1982] and Yeh and Liu [1982]. Figure 1 taken from S. Basu et al. [1988] provides a synopsis of the global distribution of scintillation occurring regions during the solar maximum and minimum periods.

Since the maximum scintillation amplitude occurs in the equatorial region, specialized reviews have appeared which deal only with this region [Aarons, 1977; Basu and Basu, 1981; 1985]. The latter two reviews discuss the impact of equatorial plasma density depletions or "bubbles" caused by the Rayleigh Taylor instability [Ossakow, 1981] on scintillation magnitude and longitudinal variability in this region. Further, data from the equatorial anomaly region have been used to study the effects of scintillations on communications and radar systems through the variability of spectral

"WORST CASE" FADING DEPTHS AT L-BAND

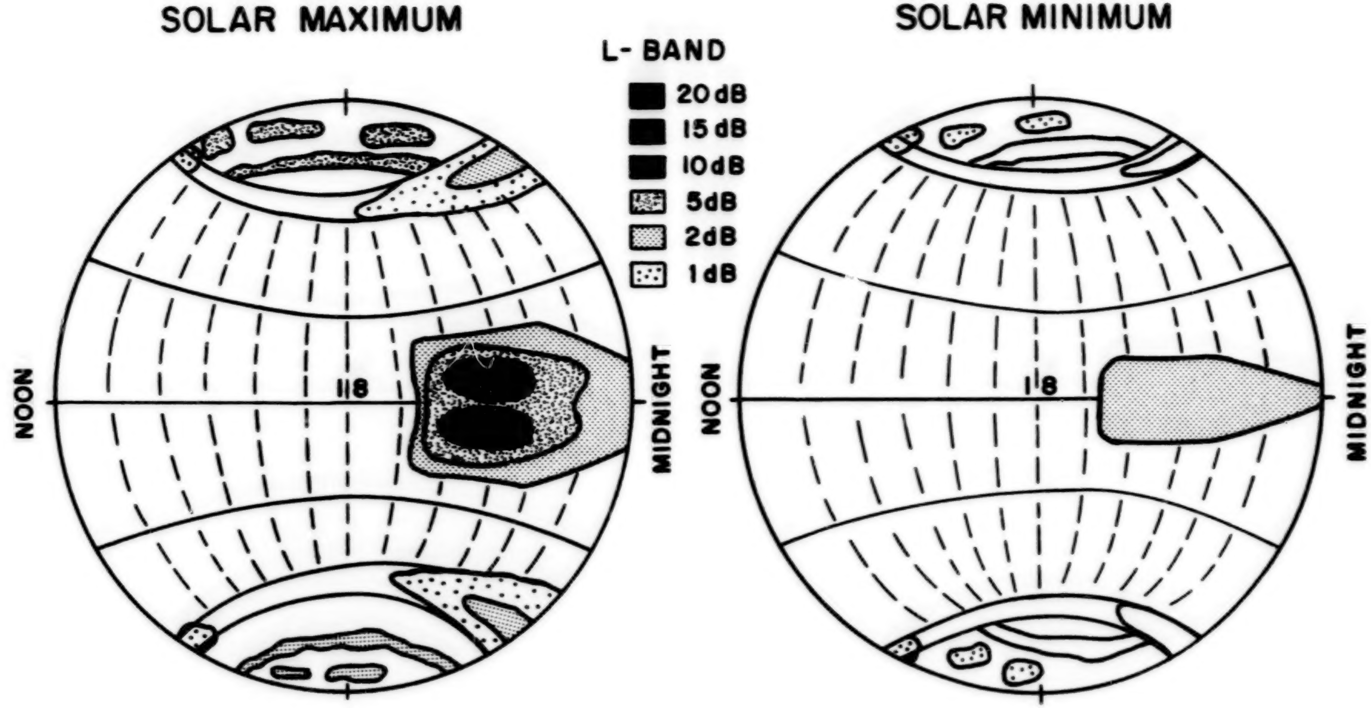


Figure 1. Global variation of scintillation fades during solar maximum and solar minimum.

shape, decorrelation time, cumulative distribution functions of signal amplitude, fade durations, the distribution of phase and intensity rates, and depolarization effects caused by diffractive scattering [Su. Basu et al., 1983; S. Basu et al., 1987; Lee et al., 1982; Franke and Liu, 1983].

At high latitudes the scintillation technique was the first to establish that the polar cap was the seat of large amplitude irregularities during solar maximum [Aarons et al.., 1981; Su. Basu et al., 1985] which could be caused by the gradient-drift instability occurring at the edges of convecting plasma density enhancements known as patches [Weber et al., 1984; Tsunoda, 1988] or by F-layer sun-aligned arcs [Weber and Buchau, 1981]. The DNA Wideband satellite did much to elucidate the sheet-like anisotropy of auroral irregularities and their morphology [Fremouw et al., 1977; Rino et al., 1978; Rino and Matthews, 1980], while the DNA HiLat satellite with its complement of scintillations and in-situ package was instrumental in identifying velocity shears as a viable source of small-scale irregularities in the ionosphere [Su. Basu et al., 1986].

Thus we find that scintillation measurements, particularly in conjunction with other diagnostics, have contributed greatly to our understanding of plasma processes in the coupled magnetosphere-ionosphere-thermosphere system. Multi-frequency scintillation studies have provided guidelines for the design of communication systems.

ACKNOWLEDGEMENT

The work at Emmanuel College was supported by AFGL Contract F19628-86-K-0038.

REFERENCES

Aarons, J., Equatorial scintillations: A review, IEEE Trans. Antennas Propagat., AP-25, 729, 1977.

Aarons, J., Global morphology of ionospheric scintillations, Proc. IEEE, 70, 360, 1982.

Aarons, J., J.P. Mullen, H.E. Whitney, A. Johnson, and E. Weber, VHF scintillation activity over polar latitudes, Geophys. Res. Lett., 8, 277, 1981.

Basu, S. and Su. Basu, Equatorial scintillations - a review, J. Atmos. Terr. Phys., 43, 473, 1981.

Basu, Su. and S. Basu, Equatorial scintillations: Advances since ISEA-6, J. Atmos. Terr. Phys., 47, 753, 1985.

Basu, S., E.M. MacKenzie, Su. Basu, E. Costa, P.F. Fougere, H.C. Carlson, Jr., and H.E. Whitney, 250 MHz/GHz scintillation parameters in the equatorial, polar and auroral environments, IEEE J. Selec. Areas Commun., SAC-5, 102, 1987.

Basu, S., E. MacKenzie, and Su. Basu, Ionospheric constraints on VHF/UHF communications links during solar maximum and minimum periods, Radio Sci., 363, 1988.

Basu, Su., S. Basu, J.P. McClure, W.B. Hanson, and H.E. Whitney, High-resolution topside in-situ data of electron densities and VHF/GHz scintillations in the equatorial region, J. Geophys. Res., 88, 403, 1983.

Basu, Su., S. Basu, E. MacKenzie, and H.E. Whitney, Morphology of phase and intensity scintillations in the auroral oval and polar cap, Radio Sci., 20, 347, 1985.

Basu, Su., S. Basu, C. Senior, D. Weimer, E. Nielsen, and P.F. Fougere, Velocity shears and sub-km scale irregularities in the nighttime auroral F-region, Geophys. Res. Lett., 13, 101, 1986.

Franke, S.J. and C.H. Liu, Observations and modeling of multi-frequency VHF and GHz scintillations in the equatorial region, J. Geophys. Res., 88,. 7075, 1983.

Fremouw, E.J., C.L. Rino, R.C. Livingston, and M.D. Cousins, A persistent subauroral scintillation enhancement observed in Alaska, Geophys. Res. Lett., 4, 539, 1977.

Fremouw, E.J., R.L. Leadabrand, R.C. Livingston, M.D. Cousins, C.L. Rino, B.C. Fair, and R.A. Long, Early results from the DNA Wideband satellite experiment - complex-signal scintillation, Radio Sci., 13, 167, 1978.

Lee, M.C., A. DasGupta, J.A. Klobuchar, S. Basu, and Su. Basu, Depolarization of VHF geostationary satellite signals near the equatorial anomaly crest, Radio Sci., 17, 399, 1982.

Ossakow, S.L., Spread-F theories - a review, J. Atmos. Terr. Phys., 43, 437, 1981.

Rino, C.L., A power-law phase screen model for ionospheric scintillation, 1, Weak scatter, Radio Sci., 14, 1135, 1979.

Rino, C.L. and S.J. Matthews, On the morphology of auroral zone radio wave scintillation, J. Geophys. Res., 85, 4139, 1980.

Rino, C.L., R.C. Livingston, and S.J. Matthews, Evidence for sheet-like auroral ionospheric irregularities, Geophys. Res. Lett., 5, 1039, 1978.

Tsunoda, R.T., High-latitude F-region irregularities: a review and synthesis, Revs. Geophysics, 26, 719, 1988.

Weber, E.J. and J. Buchau, Polar cap F layer auroras, Geophys. Res. Lett., 8, 125, 1981.

Weber, E.J., J. Buchau, J.G. Moore, J.R. Sharber, R.C. Livingston, J.D. Winningham, and B.W. Reinisch, F layer ionization patches in the polar cap, J. Geophys. Res., 89, 1683, 1984.

Whitney, H.E., J. Aarons, and C. Malik, A. proposed index for measuring ionospheric scintillations, Planet. Space Sci., 17, 1069, 1969.

Yeh, K.C. and C.H. Liu, Radiowave scintillations in the ionosphere, Proc. IEEE, 70, 324, 1982.

Chapter 5

MODERN TOTAL ELECTRON CONTENT MEASUREMENT TECHNIQUES

John A. Klobuchar

Ionospheric Physics Division
Air Force Geophysics Laboratory
Hanscom AFB, MA 01731

ABSTRACT

Since the mid-1950s measurements have been made of the total electron content, (TEC), of the earth's ionosphere by various techniques, beginning with Faraday rotation of lunar reflected VHF radio waves. With the advent of artificial earth satellites, in 1957, Faraday rotation of VHF signals from low orbit satellites became a means of determining TEC versus satellite latitude. Beginning in the mid-1960s continuous measurements of TEC at one geographic location became possible using the Faraday rotation technique, from VHF telemetry signals of opportunity transmitted from geostationary satellites. The differential Doppler technique has been used since the early 1960s to correct automatically for ionospheric range-rate errors on low orbit navigation satellites, and beginning in the mid-1970s both differential group delay and differential Doppler have been used on a high-orbit satellite navigation system to correct for first order ionospheric effects on range and range-rate. All three techniques can be used to give measurements of relative TEC, but they each have different potential problems of calibration to obtain absolute TEC from the measured values.

Because measurements of TEC have largely depended upon using radio signals transmitted from satellites of opportunity, few attempts have been made to standardize the technique, to combine TEC data from many stations to improve TEC prediction capability, or to use TEC data directly to make realistic models of the topside ionosphere. Now the potential availability of suitable radio signals from various satellites for making measurements of TEC is undergoing a transition, and will provide new opportunities for making routine measurements of this parameter.

An improved knowledge of the behavior of TEC is becoming increasingly important as a means of doing research on ionospheric behavior, especially during disturbed ionospheric conditions. In addition, the time delay effects of TEC will become even a more important source of potential error in modern trans-ionospheric positioning systems as the required accuracy of modern satellite systems inevitably increases. Realistic models and methods for correcting for the time delay effects of the TEC on such modern systems require new TEC measurement capability. This paper describes the basic measurement techniques and various methods of making such measurements.

1. INTRODUCTION

There are basically three direct methods of measuring the total electron content of the ionosphere using radio techniques. These are: 1) Faraday rotation; 2) group delay; and 3) differential carrier phase. All of these techniques have been used in some form or other in making measurements of the ionosphere, or in systems which automatically correct for ionospheric time delay effects and its rate of change. All of these techniques have disadvantages in measuring absolute values of TEC; yet, these difficulties have been overcome so that many groups have made measurements of TEC from radio signals transmitted from various satellites of opportunity. In a few cases, satellite radio beacons have been launched specifically for making measurements of TEC, as well as for amplitude and phase scintillation, but these satellites have been relatively short lived, and have generally required more complex receiving instrumentation than has been easily obtained by many potential observers. Hence the TEC data base available from these dedicated beacon satellites is very limited in its geographic coverage. However, dedicated beacon satellites have provided important new information about the behavior of TEC from those regions of the world where observations were made. Davies, [1980] has reviewed the results from the dedicated ionospheric beacon on board the ATS-6 geostationary satellite. Fremouw, et. al., [1985], described the HILAT ionospheric beacon on a low polar orbiting satellite, and Fremouw, et. al., [1978], showed the first results from the Wideband low orbit dedicated beacon satellite experiment. A variation of the differential phase method of measuring relative changes in TEC to determine absolute TEC values was proposed by Burns and Fremouw, [1970], and was first attempted on the Wideband dedicated beacon satellite.

Measurements of the TEC of the earth's ionosphere are important for several reasons. The existing models of the topside of the ionosphere are not sufficiently accurate for system's designers or operators to use to make a state-of-the-art correction for the time delay effects of the ionosphere on their systems. Most of the attention in the ionospheric modeling community, both with empirical modelers, as well as with those who construct models from physical first principles, has been in generating and improving the bottomside ionosphere to satisfy the requirements of high frequency propagation. This concentration on the bottomside ionosphere has been because of the greater number of potential users of their models in high frequency propagation.

Now, with the increasing number of people involved in satellite ranging systems requiring high precision, those who use satellites for precise time transfer, radio astronomers who require a knowledge of the effects of the ionosphere on their measurement accuracy, and other applications groups, it is becoming more important to have continuous measurements of carefully calibrated values of TEC for these many applications. The improved models of TEC will then follow, but only when newer TEC data sets become available can the modelers be assured that their efforts are an improvement over existing TEC models.

The world has approximately one hundred ionospheric sounders making routine measurements of the bottomside of the earth's ionosphere. Many of these stations report their important propagation parameters to one of the World Data Centers on an hourly, or a 24 hour summary, basis. These parameters are very important in predicting HF propagation paths on an hour to hour basis, and the 24 hour summaries are potentially very important in improving models of the bottomside ionosphere. Unfortunately, no equivalent network for making TEC measurements exists. There are perhaps no more than a twenty stations in the entire world where routine measurements of TEC are made, and only half of those stations report their data to one of the World Data Centers, either in near-real-time for improving prediction capabilities, or in summary form so that modelers can use the data to compare with their model improvement efforts.

Hopefully, the TEC measurement techniques presented in this paper will give the ionospheric research community sufficient additional information to begin to make measurements of TEC to aid in solving the problems of the day-to-day variability of the topside of the ionosphere. Also new TEC data from wide areas of the world, over sufficient time intervals, particularly during magnetic storms, and during various solar activity conditions, will be vital to ionospheric modelers in making realistic improvements to models, particularly of the topside, of the ionosphere. These improved TEC models will be of increasing importance to many trans-ionospheric propagation systems users in future years as their operational requirements for ionospheric time delay become more stringent. In this paper, first an outline of the possible methods of measuring TEC will be given, with a brief historical view of how each of these techniques has been used in the past. Then, potential methods of measuring TEC using new satellite signals of opportunity will be outlined.

2. FARADAY ROTATION

When a linearly polarized radio wave traverses the ionosphere the wave undergoes rotation of the plane of this linear polarization. At frequencies of approximately 100 MHz and higher the amount of this polarization rotation can be described by:

$$\mathrm{OMEGA} = K/f^2 * \int B \cos(\theta) * N\,dh \quad \text{(radians)} \qquad 1.$$

where the quantity inside the integral is the product of electron density times the longitudinal component of the earth's magnetic field, integrated along the radio wave path. Many ionospheric workers have used this effect, named for Michael Faraday who first observed polarization changes in an optical experiment, to make measurements of the TEC of the ionosphere.

Since the longitudinal magnetic field intensity, $B * \cos(\theta)$, changes much slower with height than the electron density of the ionosphere, equation 1, can be rewritten as:

$$\mathrm{TEC} = \frac{\mathrm{Omega} * f^2}{K * B_L} \qquad 2.$$

where $B_L = B * \cos(\theta)$ is taken at a mean ionospheric height, usually near 400 km, $K = 2.36*10^{-5}$, and TEC is $\int N\,dl$.

Typical values of polarization rotation for northern mid-latitude stations monitoring radio waves from a geostationary satellite near their station meridian are given in Figure 1, as a function of system frequency and TEC. Generally, the equivalent vertical TEC is determined by dividing the slant TEC by the secant of the zenith angle at a mean ionospheric height. The equivalent vertical TEC is the one most often used for comparison purposes among sets of TEC data, due to various slant elevation angles at which satellites at normally viewed.

At 136 MHz the one way Faraday rotation observed from a geostationary satellite can be ten, or more, 180° half turns of linear polarization change. Since the measurement is made on a single frequency, there is an n * pi ambiguity in the total amount of polarization rotation; thus, the absolute TEC cannot be determined independently using this technique. The normal method of estimating this n * pi ambiguity is to compare the relative TEC determined from the Faraday rotation with values of N_{max} obtained from an appropriately located ionosonde, during a time of night, when both the TEC and the density at the peak of the F2 region are low, and to rely on continuous polarization data for long periods, to insure the absolute calibration at other times. This method works well if careful attention is paid to obtaining continuous polarization data for long periods of time. Webb, [1973], described a method of resolving the n * pi ambiguity in 150 MHz two-way lunar reflected Faraday

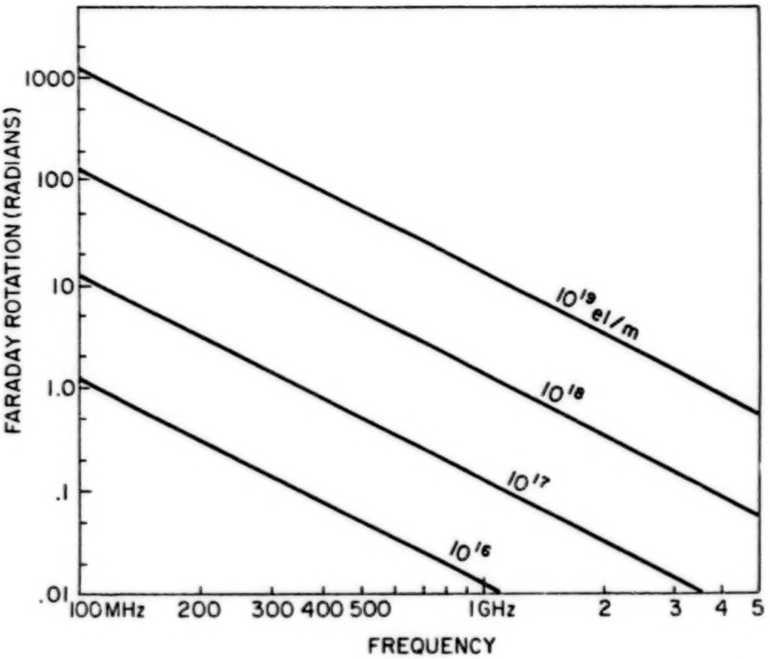


Figure 1. Faraday rotation vs frequency for various values of TEC.

rotation measurements using a fit against solar 2800 MHz radio flux.

Relatively inexpensive methods of measuring Faraday polarization changes from VHF signals transmitted from geostationary satellites of opportunity have been described by Titheridge, [1966], who used a physical rotating yagi antenna, and by Antoniadis, et. al., [1974], who developed an electronic means of generating an equivalent rotating antenna. Eis, et. al., [1977], described many aspects of determining absolute values of TEC from Faraday rotation from VHF radio waves transmitted from geostationary satellites.

The Faraday rotation technique, monitoring radio transmissions from 136 MHz telemetry beacon transmitters on geostationary satellites has been the standard technique in the past for making TEC measurements. However, modern geostationary satellites, with transponder traveling wave tube amplifiers at higher frequencies, have become so reliable that they no longer have the requirement for an independent beacon transmitter; thus, they no longer carry VHF telemetry beacon transmitters.

Effective in January 1990, the World Administrative Radio conference has re-allocated the 136-137 MHz portion of frequency spectrum to the Aeronautical Mobile Service for use on a primary basis, and has recommended that consideration be given to deleting all secondary allocations from the 136-137 MHz band. Thus, even though the geostationary satellite VHF beacon transmissions are only at the one watt power level, and are transmitted from a platform approximately 40,000 km from the earth, it appears that soon the era of making TEC measurements from VHF signals transmitted from geostationary satellites will be at an end. This need not be the case, as the approximate 1 watt transmissions from VHF telemetry signals from a space platform, at geostationary satellite distance from the earth, certainly would not interfere with the Aeronautical Mobile Service. Also, modern techniques of transmitting pseudo-random-noise encoded signals would easily allow the joint use of this portion of the spectrum. Any potential future satellite ionospheric beacon designers need not be concerned about the potential for interfering with other active radio frequency spectrum users, if modern signal transmission and processing techniques are used.

The past approximate 25 years of VHF telemetry transmissions has allowed many observers to make TEC measurements using the Faraday rotation technique. Since this era appears to be drawing to a close another technique will likely be used for future measurements of TEC.

3. DIFFERENTIAL CARRIER PHASE

As a radio signal traverses the ionosphere the phase of the carrier of the radio frequency transmission is advanced. The amount of this carrier phase advance, also called phase path decrease, can be expressed as:

$$\mathrm{Phi} = \frac{1.34 * 10^{-7}}{f} * \mathrm{TEC} \quad \text{(cycles)} \qquad 3.$$

where f is the system operating frequency in Hertz, and TEC is in units of electrons/m^2 column. In practice, the amount of this phase advance cannot readily be measured on a single frequency, unless both the transmitter and the receiver have exceptional oscillator stability and the satellite orbital characteristics are well known. Usually two, coherently derived, frequencies are required for this measurement.

If two, coherently derived, frequencies are indeed transmitted from a satellite, the differential phase shift between the two frequencies can be measured. That differential measurement is related to TEC by:

$$\text{Delta (Phi)} = \frac{1.34}{f_L}*10^{-7} *(m^2 -1)/m^2 * \text{TEC (cycles)} \qquad 4.$$

where $m = f_H/ f_L$. The U. S. Navy TRANSIT system, described by Black, [1980], has used this technique for automatic correction for first order ionospheric range-rate error correction since the system was first launched in the 1960s. Even though the U.S. Department of Defense intends to terminate its control over the TRANSIT navigation system by the mid-1990s, due to the imminent completion of the high-orbit NAVSTAR Global Positioning System, (GPS), the installed number of commercial navigation receivers is very large, and it is likely that the TRANSIT satellite navigation system will continue to operate well into the 21st century. As of mid-1989 there were over ten TRANSIT satellites in orbit. Someone desiring to construct an ionospheric monitoring system using signals from the TRANSIT navigation signals should expect to have signals available for at least a decade.

It is possible to construct a relatively inexpensive dual frequency receiver, for measuring differential carrier phase between the 150 and the 400 MHz TRANSIT satellite carriers, by converting one of the commercially available 400 MHz, single frequency, consumer navigation receivers to dual frequency operation by using the local oscillator energy from the 400 MHz receiver in a simple 150 MHz secondary channel. By taking advantage of the existing phase locked loop of the 400 MHz channel, one can divide the audio frequency by the 8/3 ratio required to directly phase compare against the audio from the 150 MHz channel. Except during times of strong amplitude and phase scintillation, a simple, dual-frequency, receiver using, as a base, an inexpensive commercially available 400 MHz receiver, should work well. If a digital output is available, which has demodulated satellite orbital elements and ID information of the Transit satellite being monitored, an automatic Transit data collection system can be assembled. Leitinger, et. al., [1975], has described a technique for combining data from Transit passes received at two, or more, sites to resolve the ambiguity in absolute TEC from the differential phase data.

4. GROUP DELAY

The additional time delay, over the free space transit time, of a signal transmitted from above the ionosphere, to a user on, or near, the earth's surface, is given by Millman, [1965], and can be expressed as:

$$\text{Delta t} = \frac{40.3}{cf^2} * \text{TEC} \quad \text{(seconds)} \qquad 5.$$

where TEC is the total number of electrons along the path from the transmitter to the receiver, c is the velocity of light in meters/sec., and f is the system operating frequency in Hertz. The TEC is the number of electrons in a unit cross section column of one square meter area along this path. Figure 2 illustrates the additional time delay encountered by a radio wave, as it traverses the ionosphere, as a function of its frequency.

By measuring the group path delay independently at two, widely spaced frequencies, the TEC along the path from satellite to receiver can be measured directly. If two system operational frequencies are chosen, f_1, and a lower

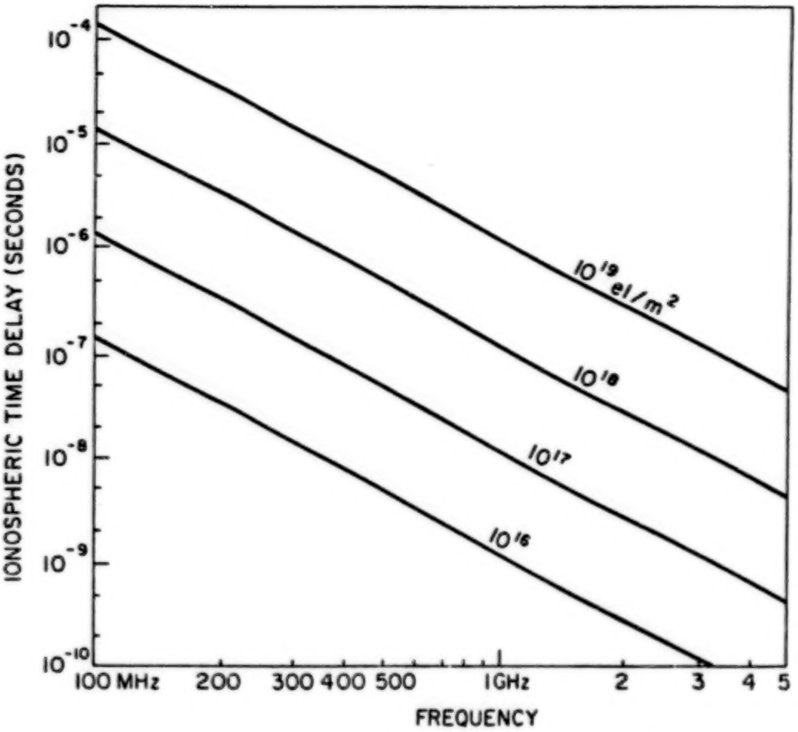


Figure 2. Time delay vs. frequency for various values of TEC.

frequency, f_2, from 5., above we obtain:

$$\mathrm{Delta}_2\ t = \frac{K}{c} * \mathrm{TEC}\ [1/f^2{}_2 - 1/f^2{}_1] \qquad 6.$$

$$\mathrm{Delta}_2\ t = \mathrm{Delta}\ t_1\ [f^2{}_1 - f^2{}_2] / f^2{}_2 \qquad 7.$$

where Delta t_1 is the ionospheric time delay at f_1.
For a direct measure of absolute TEC from the 10.23 MHz modulation phase delay at L_2 minus that at the higher frequency, L_1, we obtain:

$$\mathrm{TEC} = 2.852 * 10^{16} * \mathrm{Delta}_2\ t, \qquad 8.$$

A more detailed derivation of the equations for differential group delay and differential carrier phase advance can be found in Klobuchar [1985], section 10.8.

The Global Positioning System, (GPS), described by Denaro, [1981], and Parkinson and Gilbert, [1983], will consist of 21 satellites in 55 degree inclination, 12 hour sidereal orbits, such that a minimum number of four satellites will be visible to any user on, or near, the earth's surface, at all times. The GPS satellites transmit dual frequency, coherently derived, L-band signals, to automatically correct the first order ionospheric range and range-rate errors for navigation system users. The transmitted signals are the 120th and 154th harmonics of a standard 10.23 MHz frequency. These signals can also be used by ionospheric researchers to measure differential carrier phase and differential group delay, to make precise, absolute calibrated measurements of TEC.

The dual frequency, L-band signals transmitted from the GPS satellites are modulated with a pseudo-random noise code, having selective availability, so that the casual user cannot easily make ionospheric measurements from these signals. However, several groups, including MacDoran, et. al., [1984], have worked on either completely, or partially code-free, GPS receivers, for geodetic measurements, some with the potential for making ionospheric measurements. French and Gardner, [1986] discuss despreading a spread-spectrum signal without knowledge of the code, and at least two such receivers are presently close to commercial availability, [Osborne, (private communication), Allen, (private communication)]. In addition, the Jet Propulsion Laboratory is currently developing a dual frequency GPS receiver, called the ROGUE, Thomas, [1988], which will be capable of making ionospheric range corrections, for precise measurements required in NASA's TOPEX program. These new GPS code-free equipment developments bear close watching, as the future of modern TEC measurements will be greatly enhanced by successful, development of inexpensive receivers to make multi-direction differential group delay measurements from the numerous GPS satellites.

5. A MODERN GEOPHYSICAL BEACON CONCEPT

A major disadvantage in measuring TEC at individual ground stations, from radio signals of opportunity transmitted from satellites, is the need to have careful, uniform calibrations at all the stations, as well as timely availability of the data from each station. A concept has been proposed by Klobuchar and Hicks, [1989], in which a geostationary satellite would have an on-board beacon transponder which would simply receive signals from a potentially large number of low-power beacon transmitters on the ground. The signals received at the satellite would be re-transmitted to the earth on an S-band transponder frequency.

The low power ground transmitters would have both VHF and UHF pseudo random noise signals, separately coded for each station, and could carry low data rate geophysical information modulated onto the signals. For instance, meteorological or seismic information could be transmitted from each low power ground station, to a central data collection facility, via the geostationary satellite geophysical transponder. The complete band received at the satellite would simply be translated to one of the standard TV satellite transmit channels and could be received by a modest size antenna and receiving system similar to that currently used by consumer receivers of direct satellite TV transmissions.

The centrally located, S-Band ground receiver would decode the modulation from each beacon transmitter, demodulate the low data rate geophysical information, and make the necessary differential carrier phase, differential group delay and Faraday rotation calculation for each ground beacon transmitter. All the ionospheric and other geophysical information from potentially over one hundred ground locations would then be available at one central station for immediate use in ionospheric and other geophysical predictions.

It is even possible that, through the clever use of signal encoding, such beacon transponded signals would not interfere with the normal use of a 6 MHz S-Band satellite transponder channel. Thus, a typical satellite transponder channel could jointly transmit a standard television signal as well as the geophysical information, received at the satellite, from over one hundred ground locations within view of the geostationary satellite.

6. IONOSPHERIC TOMOGRAPHY

The measured quantity from either the Faraday rotation, the differential Doppler, or the differential group delay, technique is just the one-dimensional value of TEC along a specific direction. Austen, et. al., [1988], showed that by using several TEC monitoring stations spaced below a low orbit satellite pass, one could obtain the two dimensional electron density versus height profile along the path of the satellite. Electron density profiles over a large geographic area are of potentially great importance to those who require information on the height profile of electron density, such as in high frequency propagation, in looking at the details of ionospheric variability during magnetically disturbed times, and in modeling studies, where assumptions of scale height of both the topside and the bottomside of the ionosphere shape can be verified using the tomography technique. Though only limited tests of this technique have been done thus far, it holds great promise as a modern, inexpensive method of determining electron density profile information over a large geographic region.

7. A SOFTWARE RECEIVER

With the ever increasing speed of modern digital computers, and their decreasing costs, a major portion of analog receivers and data collection can be performed, in a more general way, by the use of digital techniques. Chiralo and Spalla [1989] described a method in which, after suitable analog RF amplification and prefiltering is done to prevent aliasing, the signal can be digitally processed in a number of ways. These include mixing, additional filtering, phase and amplitude detection, phase comparison, spectral analysis, and other sampling and statistical computations. They illustrated a software approach to a VHF polarimeter, pointing out the relative low cost of duplicating such a receiver after the investment in initial development is completed.

The future of software receiver techniques in TEC measurements, from satellite signals of opportunity, is large. The potential advantages of a software receiver, over an analog system, include low cost, assurance of reproducible calibrations among the receivers, and great versatility of data collection and reduction.

8. TEC DATA REQUIREMENTS

The requirements for future measurements of TEC include the following:

1. Data taking from many locations throughout the world
2. Standardized digital format
3. Absolute calibration, rather than only relative measurements
4. Fully automated operation
5. Continuous data from each station
6. Availability of different data collection rates
7. Centralized, remote data collection at World Data Centers
8. Low or moderate cost
9. Data from many directions from each station

While some of these requirements seem contradictory, namely the low cost versus automatic operation, continuous data from many directions, in this age of modern, low cost computational facilities, and ever decreasing electronic component costs, it is possible to construct such a system. The dual frequency L-band signals of opportunity from the GPS satellites represent the best opportunity for future TEC measurements, providing inexpensive receiving systems are designed and fielded.

9. CONCLUSIONS

The routine measurement of the earth's total electron content is very important in studies of the physics of ionospheric F2 region behavior, especially as more interest focuses on the dynamics of the topside ionosphere. Coupled with nearby measurements of the bottomside ionosphere with modern ionosondes, the TEC can be used to yield a topside ionospheric profile shape factor. Normally several satellites can be viewed in different directions at once, thus yielding important information on the gradients in the ionosphere within approximately plus and minus ten degrees of earth center angle around the observing station.

TEC measurements also are becoming increasingly important in corrections to advanced ranging systems, precise time transfer by satellite, and in correction for radio astronomical measurements. Improved models of the topside ionosphere must rely on actual TEC data for their construction and verification. The projected availability of the constellation of 21 GPS satellites of opportunity, which can be used to make absolute TEC measurements, as well as precise relative TEC measurements, heralds the beginning of a new era in trans-ionospheric propagation measurement capability. The new technique of tomography, applied to constructing electron density profiles from TEC information from a number of stations, also shows great promise. Finally, newer generation software receivers, relying heavily of the use of computers, will add an important source of versatility to the existing measurement capability.

REFERENCES

Antoniadis, D. A.[1974], IEEE Trans. on Aerospace and Electronic Systems., AES-10, 510.

Austen, J. R., S. J. Franke, and C. H. Liu, [1988], Radio Science, 23, 299.

Black, H. D., [1980], Philos. Trans. R. Soc. Lond., A 294, 217.

Burns, A. A., and Fremouw, E. J., [1970] IEEE Trans. on Ant. & Prop. , AP-18.

Ciraolo, L. and P. Spalla, [1989], in INVESTIGATION OF THE IONOSPHERE BY MEANS OF BEACON SATELLITE MEASUREMENT, published by International Academic Publishers, Xizhimenwai Dajei, Beijing Exhibition Centre, Beijing 100044, China, distributed by Pergamon Press, Oxford, U. K.

Davies, K., [1980], Space Sci. Rev., 25, 357.

Denaro, R. P., [1981], IEEE Spectrum, 18, 35.

Evans, J. V., [1977] Reviews of Geophys. & Space Phys., 15, 325.

Fremouw, E. J., R. L. Leadabrand, R. C. Livingston, M. D. Cousins, C. L. Rino, B. C. Fair, and R. A. Long, [1978], Radio Science, 13, 167.

Fremouw, E. J., H. C. Carlson, T. A. Potemra, P. F. Bythrow, C. L. Rino, J. F. Vickrey, R. L. Livingston, R. E. Huffman, C. I. Meng, D. A. Hardy, F. J. Rich, R. A. Heelis, W. B. Hanson, and L. A. Wittwer, [1985], Radio Science, 20, 416.

French, C. A., and W. A. Gardner, [1986], IEEE Trans. on Comm., COM-34, 404.

Klobuchar, J. A. [1985] HANDBOOK OF GEOPHYSICS AND THE SPACE ENVIRONMENT, A. S. Jursa, Editor, Document Accession Nr. ADA167000, available from National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161.

Klobuchar, J. A., and P. A. Hicks, [1989], in INVESTIGATION OF THE IONOSPHERE BY MEANS OF BEACON SATELLITE MEASUREMENT, published by International Academic Publishers, Xizhimenwai Dajei, Beijing Exhibition Centre, Beijing 100044, China, distributed by Pergamon Press, Oxford, U. K.

Leitinger, R., G. Schmidt, & A. Tauriainen, [1975], J. Geophys., 41, 201.

MacDoran, P. F., J. H. Whitcomb, and R. B. Miller, [1984], Sea Technology, 25, No. 10, 10.

Millman, G. H., [1965], in MODERN RADAR, Ed. by R. S. Berkowitz, John Wiley & Sons, New York.

Parkinson, B. W. and Gilbert, S. W., [1983], Proc. IEEE, 71, 1177.

Thomas, J. B., [1988], JPL Publication 88-15, June 1.

Titheridge, J. E., [1972], Planet. Space Sci., 20, 353.

Titheridge, J. E., [1966], J. Atmos. Terr. Phys., 73, 1135.

Webb, H. D., [1973], Journ. Geophys. Res., 78, 1214.

Chapter 6

RAYLEIGH AND SODIUM LIDAR TECHNIQUES FOR MEASURING MIDDLE ATMOSPHERE DENSITY, TEMPERATURE AND WIND PERTURBATIONS AND THEIR SPECTRA

Chester S. Gardner, Daniel C. Senft, Timothy J. Beatty,
Richard E. Bills and Chris A. Hostetler

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
1406 West Green Street - Everitt Laboratory
Urbana, IL 61801

ABSTRACT

We review techniques for measuring middle atmosphere density, temperature and wind perturbations using Rayleigh and Na lidar systems. We restrict our attention to the problem of measuring the mean-square perturbations of these parameters and to the problem of computing relative density and wind spectra. The effects of photon noise on the accuracies of these measurements and the influence of laser power and telescope area on lidar system performance are discussed in detail. Examples of Rayleigh and Na lidar measurements of atmosphere density and temperature profiles and the vertical wavenumber and temporal frequency spectra of relative density and wind perturbations are presented.

I. INTRODUCTION

The middle atmosphere is a region of complex photochemical and dynamic interaction. Stretching from roughly 30 to 100 km altitude, it is the transition region between the stratosphere and thermosphere and is perhaps the least understood region of the earth's atmosphere. Inaccessible to balloons and satellites, observations of this region must be made by direct *in situ* rocket measurements or by satellite borne or groundbased remote sensing techniques. The middle atmosphere is characterized by a variety of interesting and important features. For example, the temperature minimum at the mesopause near 90 km is the coldest part of the atmosphere. The ionosphere D-region is part of the mesosphere and near 100 km is the turbopause, above which eddy mixing is absent. A variety of layered structures are found in the mesosphere including noctilucent clouds, the polar mesospheric dust layer, the airglow layers and the alkali metal layers of sodium, lithium and potassium.

The density and temperature structure of the middle atmosphere is influenced to a large extent by the effects of gravity waves and tides. Internal gravity waves are believed to be generated primarily by disturbances in the lower atmosphere. Tropospheric convection, storms, wind shears, orographic effects such as airflow over mountains, and even large-scale ocean swells are thought to be common gravity wave sources. These waves propagate through the stratosphere and mesosphere and dissipate their energy near the mesopause making important contributions to the momentum and turbulence budget in this region of the atmosphere. Although the gravity wave wind amplitudes are typically very small in the lower atmosphere, in the absence of dissipation, the velocities grow exponentially with altitude and can approach values as high as 50 m/s at mesopause heights. At these altitudes the waves can change the atmospheric density by as much as 10-20 percent over time periods as short as a few hours. When the amplitudes become too large, instabilities cause the waves to break and dissipate their energy creating turbulence.

The absorption of infrared radiation by CO_2 and water vapor in the troposphere and the absorption of UV radiation by ozone in the stratosphere and mesosphere generate atmospheric tides that also propagate through the middle atmosphere. At low- and mid-latitudes, the diurnal (24 h period) and semidiurnal (12 h period) tides appear to have the greatest influence. It is now widely recognized that both gravity waves and tides play a major role in determining the large-scale circulation and structure of the middle atmosphere [Lindzen, 1981; Fritts, 1984]. Because of their significant influence on density and temperature, the temporal and spatial characteristics of both gravity waves and tides can be studied with Rayleigh and Na lidar techniques.

In this paper we review the techniques for measuring atmospheric density, temperature and wind perturbations using Rayleigh and Na lidar systems. We restrict our attention to the problem of measuring the mean-square perturbations of these parameters and to the problem of computing the density and wind spectra. Several excellent studies of monochromatic gravity waves and tides using lidar techniques have already been published [e.g., Chanin and Hauchecorne, 1981; Batista et al, 1985; Gardner and Voelz, 1987; Kwon et al., 1987; Gardner et al., 1989]. For this reason we concentrate on the effects of the quasi-random waves which are best characterized in terms of spectra and mean-square variations.

The system configurations for typical Rayleigh and Na lidars are briefly described in Section II. The lidar equation and predicted system performance are discussed in Section III and the atmospheric density response to wind perturbations is analyzed in Section IV. The techniques for calculating the mean-square density and horizontal wind perturbations and the associated spatial and temporal spectra are discussed in detail in Section V. The effects of photon noise on the measurements of these parameters are also evaluated in this section. The techniques for calculating temperature profiles from lidar observations are described in Section VI. Practical considerations in processing Rayleigh and Na lidar data are discussed in Section VII. And finally, in Section VIII, example Rayleigh and Na lidar observations of atmospheric density, temperature and wind variations are presented.

This paper was written to provide a reasonably complete description of how to process lidar data for dynamics and temperature studies and how to estimate the accuracies of the measured atmospheric parameters. Some of the sections include considerable mathematical detail which may not be of interest to every reader. For those individuals interested in obtaining a general assessment of the capabilities of both Rayleigh and Na lidars, we recommend reading Sections II and VIII. For those readers interested in understanding the effects of photon noise on the accuracies of the measurements, we recommend reading Sections III, V and VI.

II. SYSTEM CONFIGURATION

The most common lidar configuration is a monostatic system in which the laser beam is either projected through the receiving telescope or propagates parallel to the optical axis of the telescope. Profile measurements are accomplished by pulsing the laser and range gating the detector. Less common is the bistatic configuration in which the laser and receiving telescope are separated by distances of up to 20 km. Profiling is accomplished by either scanning the telescope along the laser beam or by scanning the laser along the telescope field-of-view. Some of the earliest lidar measurements were conducted by Elterman in the early 1950s. He used a bistatic configuration consisting of a mechanically modulated searchlight and a scanning telescope to measure density and temperature profiles from 10 to 60 km altitude [Elterman, 1951; 1953; 1954]. By modulating the searchlight Elterman was able to discriminate between the scattered signal and the background noise from the stars and moon. Here, we will restrict our discussion to the design of monostatic lidars because they are the most widely used.

The basic design of a monostatic resonance fluorescence lidar system is very similar to the design of aerosol, Rayleigh and Raman lidars. Figure 1 is a block diagram of a typical Na lidar. The pulsed laser is tuned to the D_2 resonant absorption line of Na at a wavelength near 589 nm which is in the yellow-orange region of the visible spectrum. Most existing Na lidars employ dye lasers that are pumped by flashlamps, excimer lasers or Nd:YAG lasers. A small fraction (<1 percent) of the output beam is directed to wavelength and energy monitors and to a pulse detector, which triggers the data acquisiton system. The laser pulse propagates up to the Na layer where it is resonantly scattered by neutral Na atoms. The backscattered photons are collected by the telescope and focused onto a photomultiplier tube (PMT). Because the backscattered signal level is very low, the PMT must operate in the photon counting mode even when the lidar system is operated during the daytime. The field stop iris and interference filter in the receiving telescope are designed to reject much of the background light from the stars, moon, and sun. The PMT

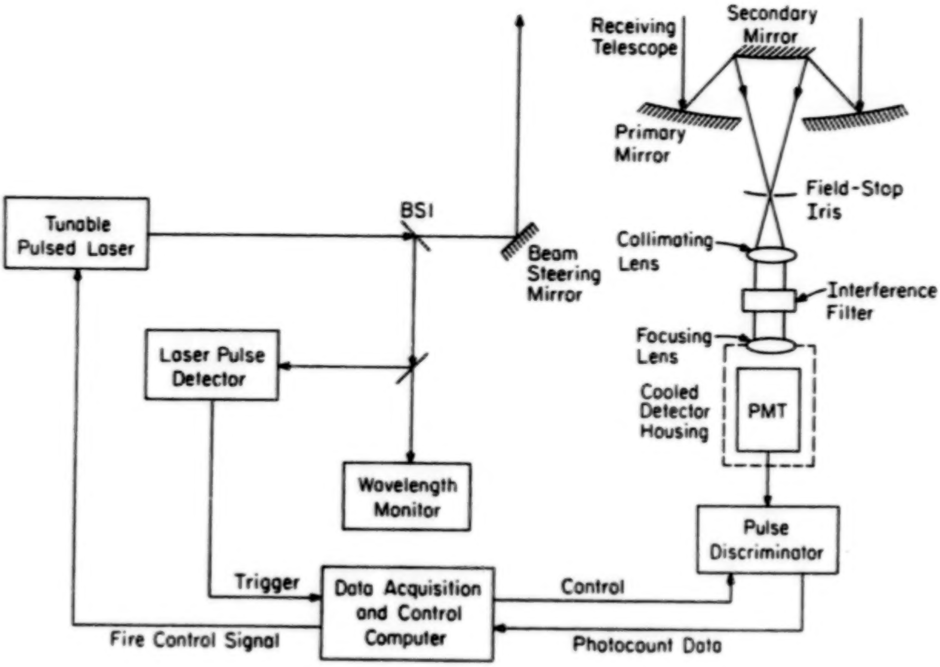


Figure 1 Block diagram of a typical Na resonance fluorescence lidar system.

signal is processed by a pulse discriminator and range-gated counter. The pulse discriminator converts the low-level (few tenths of a volt) PMT pulses to logic level pulses which can then be counted using conventional high-speed digital circuitry. The raw lidar data consists of a series of photon counts corresponding to consecutive range bins. Because the signal from a single laser pulse is not sufficient to construct a complete Na density profile, the photon counts from many pulses are accumulated over time periods ranging from tens of seconds to several minutes depending on the desired measurement accuracy.

A typical photocount profile is plotted in Fig. 2. This profile was obtained at Arecibo Observatory, Puerto Rico (18°N, 67°W) using the University of Illinois at Urbana-Champaign (UIUC) CEDAR Na lidar. The lidar operated with a 250 ns receiver range-gate which corresponds to a range bin length of 37.5 m. The profile consists of the accumulated photocounts from 4800 laser pulses obtained during an integration period of 25 s. The resonant scattering from the Na layer between about 80 and 105 km is clearly evident. The wavelike structure in the layer is caused by the wind perturbations associated with low-frequency internal gravity waves. The nonzero count level above 110 km is caused primarily by background noise from scattered moonlight and starlight.

The high photocount levels below 60 km result from Rayleigh (i.e., molecular) and Mie scattering by air molecules and aerosols. To prevent overloading by the very strong scattering in the troposphere, the gain of the PMT must be reduced by about 25 dB during the first 50 μs after the laser has fired. The rapid increase in signal level at about 8 km altitude corresponds to the point where the PMT gain is switched to its maximum value. Gain switching can be accomplished either electronically by controlling the PMT gain or mechanically by placing a high-speed rotating shutter wheel near the field stop iris in the telescope. In the UIUC CEDAR lidar system, the gain is switched by controlling the voltage on the first dynode of the PMT.

Because the laser beam propagated beside the telescope, the beam was only partially visible within the telescope field-of-view (FOV) at the lower altitudes. In the absence of dust or aerosol layers, the molecular backscattered signal increases with altitude as the overlap between the laser beam and telescope FOV increases. Eventually the $1/z^2$ (z = altitude) losses and decreasing atmospheric density dominate, the molecular signal level reaches a maximum and then decreases with increasing altitude. For the data plotted in Fig. 2, this occurs at an altitude of about 3 km.

In the troposphere and lower stratosphere, the backscattered signal is often enhanced by Mie scattering from clouds and aerosol layers. Enhanced scattering from aerosol layers at 1, 2 and 5 km can be seen in Fig. 2(b). During the past decade, observations with groundbased and airborne aerosol lidars have provided important information about the dispersion and global distribution of volcanic aerosols [e.g., McCormick et al., 1984; Reiter and Jager, 1986] and about the life cycle and distribution of polar stratospheric clouds (PSC) [e.g., Kent et al., 1986; Iwasaka et al., 1986]. Volcanic aerosols can have a significant influence on climate, while PSCs play a major role in the springtime depletion of ozone over both polar caps. Much of the existing observational data of these phenomena, particularly volcanic aerosols, were obtained with lidars.

Above approximately 25 km and below the Na layer the scattering is purely molecular so that the signal level is proportional to atmospheric density. Rayleigh lidars are designed specifically to measure the molecular scattered signal which can then be used to infer atmospheric density and temperature. The Na lidar Rayleigh signal between 25 and 70 km can be used to study the atmospheric temperature and density profile. The use of high power lasers with the Rayleigh technique was pioneered by Chanin and Hauchecorne [1981]. The data analysis approach is very similar to that employed by Elterman in the early 1950s to measure stratospheric temperatures and is discussed in more detail in Section VI.

III. LIDAR EQUATION AND SYSTEM PERFORMANCE

The theoretical performance of any lidar system is governed by the lidar equation. The expected received photon count is equal to the product of the system efficiency, number of transmitted photons, probability that a transmitted photon is scattered and probability that a scattered photon is received. For Na resonance fluorescence systems the lidar equation is given approximately by [Gardner et al., 1986]

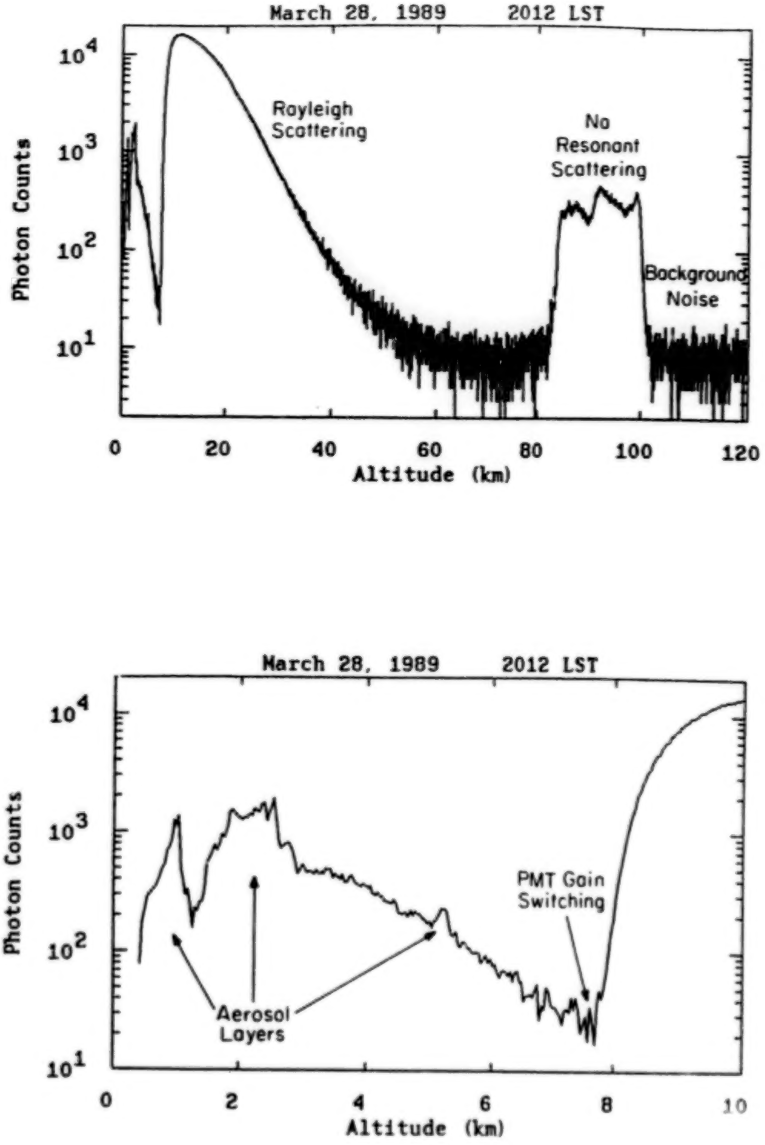


Figure 2 Lidar photon count profile obtained on the night of March 28, 1989 at Arecibo Observatory, Puerto Rico with the UIUC CEDAR Na lidar. The profile was collected by integrating the backscattered signal from 4800 laser shots during a period of 25s starting at 2012 LST. The vertical resolution of the data is 37.5 m. a) Complete profile, b) Expanded low-altitude profile.

$$N_s(z) = \left(\eta T_a^2\right)\left(\frac{P_L \Delta t}{hc/\lambda_L}\right)\left(\sigma_{eff}\, \rho_s(z)\, \Delta z\right)\left(\frac{A_R}{4\pi z^2}\right) + N_B R_L \Delta t \qquad (1)$$

where

$N_s(z)$ = expected number of Na and background photons detected in the range interval $(z - \Delta z/2,\ z + \Delta z/2)$
$\rho_s(z)$ = Na density at range z (m^{-3})
N_B = expected photocount per range bin per pulse due to background noise and dark counts
σ_{eff} = effective Na backscatter cross section (m^2)
Δz = receiver range bin length (m)
A_R = receiving telescope aperture area (m^2)
λ_L = optical wavelength of laser (0.589×10^{-6} m)
h = Planck's constant ($6.63 \times 10^{-34}\ Js^{-1}$)
c = velocity of light ($3 \times 10^8\ ms^{-1}$)
P_L = laser power (w)
T_a = one-way transmittance of the lower atmosphere (z < 80 km)
R_L = laser pulse rate (s^{-1})
Δt = integration time (s)
η = lidar efficiency

Equation (1) was derived by assuming that extinction of the laser beam as it propagates through the layer is negligible [Simonich and Clemesha, 1983] and that the power densities of the beam within the layer are small enough to ignore the effects of stimulated emission from the excited Na atoms [Welsh and Gardner, 1989]. In practice both of these effects can be significant and should be taken into account when processing Na density data.

The first factor in parenthesis in (1) is the overall system efficiency. The lidar efficiency η includes the optical efficiencies of the beam splitters, mirrors, lenses,and optical filters in the laser transmitter and receiving telescope (Fig. 1) as well as the quantum efficiency of the PMT. η is dominated by the optical filter transmittance and the detector quantum efficiency. At the Na D_2 wavelength, PMTs with GaAs photocathodes have quantum efficiencies between 10 and 20 percent. Wideband optical interference filters with bandwidths greater than 5 Å FWHM typically have transmittances between 30 and 80 percent. Narrowband filters (< 5 Å), which are required for daytime observations, have transmittances of 10 percent or less. Thus, the system efficiencies for well-designed Na lidars range between 1 and 8 percent, with smaller values corresponding to narrowband systems which can operate during day or night. The atmospheric transmittance T_a depends on the wavelength and clarity of the lower atmosphere. Absorption by ozone at near UV wavelengths and by water vapor and carbon dioxide at near IR wavelengths can significantly reduce the atmospheric transmittance. However, under excellent seeing conditions at the Na wavelength, T_a is approximately 80 percent for zenith observations.

The second factor in (1) is the number of transmitted photons. The factor $P_L \Delta t$ is the total laser energy and hc/λ_L is the photon energy. The third factor in parenthesis is the probability that a transmitted photon is scattered by a Na atom. The scattering probability is equal to the product of the effective Na backscatter cross section (σ_{eff}), Na density ($\rho_s(z)$), and the scattering layer thickness (Δz). The fourth factor is the probability that a scattered photon is collected by the receiving telescope. For isotropic scattering, this probability is equal to the telescope area (A_R) divided by the area of a sphere of radius z (i.e., $4\pi z^2$). z is the distance from the scatterer to the telescope. For convenience, effects caused by scattering anisotropies are included in the scattering cross section (σ_{eff}).

The backscatter cross section depends on the laser wavelength and linewidth and on the velocity and temperature of the scatterer. The Na D_2 resonance line is actually a Doppler broadened doublet composed of 6 lines. Figure 3 is a plot of the Na resonance line as a function of wavelength for several values of temperature. As the temperature decreases the minimum between the D_{2a} and D_{2b} lines decreases and the peak intensities of the D_{2a} and D_{2b} lines increase. The backscatter cross section also depends upon the laser linewidth. The maximum value, which is

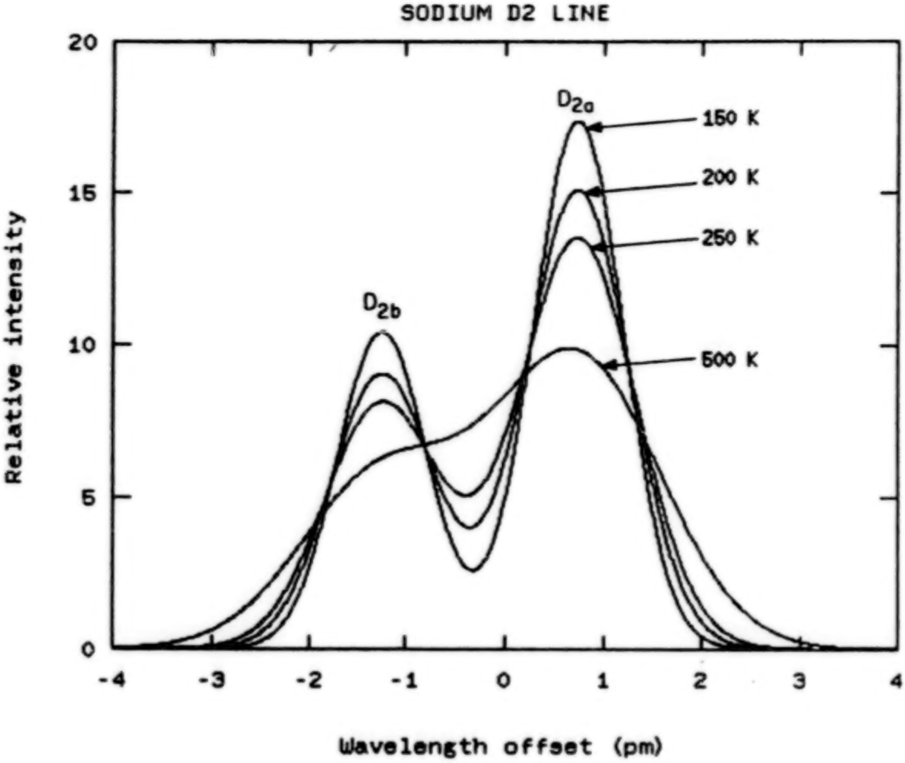


Figure 3 Thermally broadened hyperfine structure of the Na D_2 resonance line.

achieved by tuning a narrowband laser (FWHM laser linewidth < 0.5 pm) to the D_{2a} wavelength, is approximately 9.6×10^{-16} m^2 for $T \sim 200$ k.

The Na density can be estimated from the photocount data by subtracting the measured background photocount and then solving for $\rho_s(z)$ in (1)

$$\rho_s(z) = \frac{4\pi z^2 hc/\lambda_L}{\eta T_a^2 \sigma_{eff} P_L A_R \Delta z \Delta t}\left[N_s(z) - \hat{N}_B R_L \Delta t\right] \tag{2}$$

where $\hat{N}_B$ is the estimated photocount per range bin per laser pulse due to background noise and dark counts. Accurate estimates of the background count can be obtained by averaging counts in the range bins above the Na layer.

Because the atmospheric transmittance depends on the clarity of the lower atmosphere and is not easy to measure, the absolute Na density is usually computed by normalizing the Na photocounts by the Rayleigh photocount at an altitude free of aerosols, typically 30-35 km. Aerosol-free regions are relatively easy to identify because the enhanced scattering by aerosols are usually evident in the photocount profiles. Several studies using multi-wavelength lidars have shown that aerosol layers are rare above 35 km and virtually nonexistent above 40 km. The expected photocount at Rayleigh scattering altitudes is given by

$$N_R(z) = \left(\eta T_a^2\right)\left(\frac{P_L \Delta t}{hc/\lambda_L}\right)\left(\sigma_R\, \rho_a(z)\, \Delta z\right)\left(\frac{A_R}{4\pi z^2}\right) + N_B R_L \Delta t \tag{3}$$

where

$\rho_a(z)$ = atmospheric density at altitude z (m^{-3})
σ_R = Rayleigh backscatter cross secton (m^2).

At the Na D_2 wavelength $\sigma_R \rho_a$ can be expressed in terms of the atmospheric pressure and temperature

$$\sigma_R \rho_a(z) = 3.539 \times 10^{-6}\, P(z)/T(z) \tag{4}$$

where $\sigma_R \rho_a$ has units of m^{-1} and

$P(z)$ = atmospheric pressure at altitude z (mbar)
$T(z)$ = atmosphere temperature at altitude z (K).

If z_R is the normalizing altitude, then by using (3) and (4), the Na density can be written as

$$\rho_s(z) = \frac{z^2 \sigma_R \rho_a(z_R)}{z_R^2 \sigma_{eff}} \, \frac{\left[N_s(z) - \hat{N}_B R_L \Delta t\right]}{\left[N_R(z_R) - \hat{N}_B R_L \Delta t\right]} . \tag{5}$$

The accuracy of the measured Rayleigh photocount can be improved by averaging several range bins above and below z_R. However, the atmospheric density decreases approximately exponentially with altitude and to avoid unwanted biases, this characteristic must be taken into account when the average Rayleigh photocount is computed [Gardner et al., 1986].

The required signal level depends on the desired accuracy of the atmospheric parameter being measured. Usually the measurement accuracy can be expressed in terms of the total photocount N_s comprising the Na profile. At mid-latitudes the Na column abundance C_0 ranges from a summer minimum of 3×10^9 cm^{-2} to a winter maximum of about 10^{10} cm^{-2}. The yearly average value at Urbana, IL (40°N) is approximately 5×10^9 cm^{-2} [Gardner et al., 1986]. The Na layer centroid height z_o exhibits a semi-annual oscillation at Urbana with a yearly mean of approximately 92 km. If we use these values in (1), integrate over the altitude of the Na layer and assume $\sigma_{eff} = 9\times10^{-16}$ m^2, $\eta = 5$ %, and $T_a = 0.7$, then under good seeing conditions we have for the total Na photon count (N_s)

$$N_s = \eta T_a^2 \frac{\sigma_{eff} C_o}{4\pi z_o^2 hc/\lambda_L} P_L A_R \Delta t \simeq 3\times10^4\, P_L A_R \Delta t \;\;/\;\; (wm^2 s) \tag{6}$$

where the units of P_L, A_R, and Δt are respectively w, m^2, and s. If background noise effects are neglected, the measurement errors for the layer density $\rho_s(z)$, column abundance C_o, centroid height z_o, and rms thickness σ_o are given by

$$\begin{aligned} \mathrm{Std}\left[\rho_s(z)\right] &\simeq \left[C_o\rho_s(z)/\Delta z\right]^{1/2} / \sqrt{N_s} \\ &\sim 10^{-2}\left[C_o\rho_s(z)/\Delta z\right]^{1/2} / \sqrt{P_L A_R \Delta t} \end{aligned} \tag{7}$$

$$\mathrm{Std}(C_o) = C_o/\sqrt{N_s} \sim 10^{-2}\, C_o/\sqrt{P_L A_R \Delta t} \tag{8}$$

$$\mathrm{Std}(z_o) = \sigma_o/\sqrt{N_s} \sim 10^{-2}\, \sigma_o/\sqrt{P_L A_R \Delta t} \tag{9}$$

$$\mathrm{Std}(\sigma_o) = \sigma_o/\sqrt{2N_s} \sim 10^{-2}\, \sigma_o/\sqrt{2P_L A_R \Delta t}\,. \tag{10}$$

To obtain accuracies of better than 5 percent for column abundance and better than 200 m for the centroid height and rms thickness, signal levels of approximately 500 counts are required so that $P_L A_R \Delta t \sim 0.05$ wm^{2}s. This level of accuracy can be achieved with a 0.5-w laser, 0.35-m-diameter telescope, and 1-s integration period. Gravity waves typically induce Na density perturbations on the order of 0.1 to 10 percent so that signal levels on the order of 10^4-10^5 counts are required to observe wave induced density fluctuations. To measure accurately gravity wave parameters with a 0.5-w laser and 0.35-m-diameter telescope will require integration periods of 2 to 5 min. It should be noted that in practice the accuracy of the abundance measurement is closer to 10-20 percent because model values for the atmospheric pressure and temperature are typically used in (4) rather than measured values. Notice that measurement accuracies in (7)-(10) are dominated by the factor $P_L A_R \Delta t$. Increased accuracy can only be obtained by increasing the laser power, telescope area, or integration period. Consequently, the relative performance capability of a Na lidar is dictated by the performance factor $P_L A_R$.

The absolute atmospheric density can be estimated from the Rayleigh photocount data by subtracting the measured background count and then normalizing the result by the photocount at an altitude where the absolute density is determined either from an independent observation or from a model atmosphere.

$$\rho_a(z) = \frac{z^2 \rho_a(z_R)}{z_R^2} \, \frac{\left[N_R(z) - \hat{N}_B R_L \Delta t\right]}{\left[N_R(z_R) - \hat{N}_B R_L \Delta t\right]} \tag{11}$$

The accuracy of the atmospheric density deduced from the Rayleigh backscatter signal is also related to the received photon count levels. Assuming background noise is negligible, the variance of the density estimate is given by

$$\frac{\mathrm{var}\left[\rho_a(z)\right]}{\rho_a^2(z)} \simeq \frac{\mathrm{var}\left[\rho_a(z_R)\right]}{\rho_a^2(z_R)} + \frac{1}{N_R(z)} \tag{12}$$

where

$$N_R(z) = N_R(z_s)\,(z_s/z)^2\,e^{-(z-z_s)/H} \tag{13}$$

and H is the atmospheric scale height. If we assume $\eta = 5\%$, $T_a = 0.7$, $\lambda_L = 589$ nm and $z_s = 45$ km, then

$$N_R(z_s) \simeq 0.06\,P_L A_R \Delta z \Delta t \quad / \quad (\mathrm{wm^2s}) \tag{14}$$

and the relative density error caused by photon noise is

$$\frac{\mathrm{Std}\left[\rho_a(z)\right]}{\rho_a(z)} \simeq \frac{1}{\sqrt{N_R(z)}} \simeq \frac{4}{\sqrt{P_L A_R \Delta z \Delta t}}\,(z/z_s)\,e^{(z-z_s)/2H}. \tag{15}$$

In the upper stratosphere the density perturbations caused by gravity waves and tides are typically a few percent or less so that signal levels on the order of 10^5-10^6 counts are required to measure accurately the wave induced density fluctuations. A signal level of 10^5 counts at 45 km altitude can be achieved by using a 1 m diameter telescope, 10 w laser, 2 min integration period and 1 km vertical resolution.

IV. ATMOSPHERIC DENSITY RESPONSE TO WIND PERTURBATIONS

Since Rayleigh and Na lidars directly measure density profiles, we are most interested in the affects of waves on the density structure of the atmosphere. In this section we derive equations describing the atmospheric density response to wave induced wind perturbations and then use the results to express the wave parameters in terms of the lidar observations. The density response of the neutral atmosphere or of an atmospheric layer composed of a minor neutral constituent, to wave induced wind perturbations is governed by the continuity equation. By neglecting diffusion and chemical effects and by assuming that the neutral atmosphere or constituent layer is horizontally homogeneous, Gardner and Shelton [1985] have shown that the density response can be written in the form

$$\rho(\underline{p},t) = e^{-\chi}\,\rho_o(z-\xi) \tag{16}$$

where

$\rho(\underline{p},t)$ = perturbed density at position $\underline{p}$ and time t

$\rho_o(z)$ = steady-state density profile in the absence of wind perturbations

$\underline{p} = x\hat{x} + z\hat{z}$ = position vector where x is the horizontal coordinate and z is the vertical coordinate and χ and ξ are solutions to the partial differential equations

$$\frac{\partial \chi}{\partial t} = \nabla \bullet \underline{V} - \underline{V} \bullet \nabla \chi \tag{17}$$

$$\frac{\partial \xi}{\partial t} = w - \underline{V} \bullet \nabla \xi\,. \tag{18}$$

$\underline{V} = u\hat{x} + w\hat{z}$ is the atmospheric velocity field and u and w are, respectively, the horizontal and vertical velocities. Notice that χ is related to the wind divergence and ξ is related to the vertical displacement caused by the vertical wind. These two parameters should not be confused with the common notation for geopotential height and vertical gradient of potential temperature.

If the wind perturbations are sufficiently small, the second terms on the right-hand sides of (17) and (18) can be neglected [Gardner and Shelton, 1985] so that

$$\chi(\underline{\rho},t) \simeq \int_{-\infty}^{t} \nabla \cdot \underline{V}\, d\tau \tag{19}$$

$$\xi(\underline{\rho},t) \simeq \int_{-\infty}^{t} w\, d\tau \,. \tag{20}$$

In this case the density response is given by

$$\rho(\underline{\rho},t) \simeq \exp\left(-\int_{-\infty}^{t} \nabla \cdot \underline{V}\, d\tau\right)\rho_0\left(z - \int_{-\infty}^{t} w\, d\tau\right). \tag{21}$$

The density perturbations result from a multiplicative distortion due to the wind divergence and a vertical displacement caused by the vertical winds. The relative importance of these two effects depends on the spatial scale of the wave causing the wind fluctuations.

We are primarily interested in the mean-square values and spectra of the relative density perturbations ($\Delta\rho/\rho$) as a function of time and altitude

$$r(z,t) = \frac{\Delta\rho}{\rho} = \Delta \ln \rho \,. \tag{22}$$

If the unperturbed density profile is modeled as

$$\rho_0(z) = \exp\,[-f(z)]\,, \tag{23}$$

then by using (16) and (22) we have

$$r(z,t) = -\,\chi - \Delta f(z-\xi) = -\,\chi - \left[f(z-\xi) - f(z)\right]. \tag{24}$$

As a consequence of (24) we can also write

$$\rho(z,t) = \rho_0(z)e^{r(z,t)}\,. \tag{25}$$

By expanding f(z-ξ) in a Taylor series about z and retaining terms out to first order in $\underline{V}$, we obtain

$$r(z,t) \simeq -\,\chi + \frac{df(z)}{dz}\,\xi\,. \tag{26}$$

Equation (26) is a valid approximation provided

$$|\xi| << |2f'(z)/f''(z)|\,, \tag{27}$$

where the prime denotes differentiation with respect to z.

The hydrostatic equation and ideal gas law (see equations (84) and (85) in Section VI) can be used to relate the atmospheric density and temperature profiles

$$\rho_{oa}(z) = \frac{M}{RT(z)} \exp\left[- \frac{M}{R} \int_{0}^{z} \frac{g(z)}{T(z)}\, dr \right], \tag{28}$$

so that

$$f_a'(z) = \frac{T'(z)}{T(z)} + \frac{Mg(z)}{RT(z)} \tag{29}$$

$$f_a''(z) = \frac{-T''(z)}{T(z)} - \left[\frac{T'(z)}{T(z)}\right]^2 - \frac{Mg(z)}{RT(z)}\,\frac{T'(z)}{T(z)} \tag{30}$$

where

M = mean molecular weight of the atmosphere = 28.95 kg/kmol
R = universal gas constant = 8314 Nm/(kmol K)
$g(z)$ = gravitational acceleration = $g_o/(1+z/r_o)^2$
g_o = sea level gravitational acceleration = 9.807 m/s^2
r_o = mean earth radius = 6357 km.

In the upper stratosphere and lower mesosphere the temperature profile is approximately parabolic with a maximum near the stratopause

$$T(z) \simeq T_s\left[1 - \left(\frac{z-z_s}{h_s}\right)^2\right] \tag{31}$$

where

T_s = stratopause temperature (~ 250K)
z_s = stratopause height (~ 45 km)
h_s = stratopause width (~ 30 km).

By substituting (31) into (29) and (30) we obtain

$$f_a'(z) = \frac{1}{H} - \frac{2(z-z_s)}{h_s^2} \tag{32}$$

$$f_a''(z) = \frac{-2}{h_s^2} + \frac{4(z-z_s)^2}{h_s^4} - \frac{2(z-z_s)}{Hh_s^2} \tag{33}$$

where $H^{-1} = \frac{Mg}{RT_s}$ is the atmospheric scale height at the stratopause. H is approximately 7 km in the upper stratosphere. Equation (27) is easily satisfied in the upper stratosphere and lower mesosphere because the vertical displacements caused by gravity wave winds are typically a few hundred meters or less while $|2f_a'/f_a''|$ is greater than 100 km. Thus, for atmospheric perturbations (26) becomes

$$r_a(z,t) \simeq -\chi + \frac{1}{H}\left[1 - 2H(z-z_s)/h_s^2\right]\xi\,. \tag{34}$$

It has been shown that a reasonably accurate model for the unperturbed Na layer is a Gaussian profile [Gardner and Voelz, 1987]

$$\rho_{os}(z) = \frac{C_o}{\sqrt{2\pi}\,\sigma_o} \exp\left[-\frac{(z-z_o)^2}{2\sigma_o^2}\right] \tag{35}$$

where

C_o = Na column abundance
z_o = layer centroid height
σ_o = rms layer width.

In this case $f_s(z)$ for the Na layer is quadratic so that (26) becomes

$$r_s(z,t) = -\chi + \frac{(z-z_o)}{\sigma_o^2}\xi\,. \tag{36}$$

When the density perturbations are caused by gravity waves, it is possible to simplify (34) and (36). If the vertical wavelength λ_z and intrinsic frequency ω satisfy the conditions $\lambda_z << 4\pi H$ and $\omega << N$ where N is the buoyancy frequency, the gravity wave polarization and dispension relations can be used to show that [Gardner and Shelton, 1985]

$$\xi = \gamma H \chi\,. \tag{37}$$

γ is the ratio of specific heats. In this case the relative density perturbations can be written as

$$r_a(z,t) = \frac{(\gamma-1)}{\gamma H}\left[1 - \frac{2\gamma H\,(z-z_s)}{h_s^2}\right]\xi \quad \text{(atmosphere)} \tag{38}$$

$$r_s(z,t) = \frac{-1}{\gamma H}\left[1 - \frac{\gamma H\,(z-z_o)}{\sigma_o^2}\right]\xi\,. \quad \text{(Na layer)} \tag{39}$$

It is convenient to express the relative Na density perturbations in terms of the relative atmospheric density perturbations. Equation (38) is valid for the region near the stratopause. A similar expression, valid near the mesopause, can be derived by noting that the temperature profile is also approximately parabolic in the mesopause region

$$T(z) \simeq T_m\left[1 + \left(\frac{z-z_m}{h_m}\right)^2\right] \tag{40}$$

where

T_m = mesopause temperature
z_m = mesopause height (~ 90 km)
h_m = mesopause width (~ 30 km).

As a consequence of (26), (29) and (40) we have

$$r_a(z,t) = \frac{(\gamma-1)}{\gamma H}\left[1 + \frac{2\gamma H\,(z-z_m)}{h_m^2}\right]\xi\,. \tag{41}$$

The Na layer is centered approximately at the mesopause and the full-width at the e^{-2} points is only about 20 km. Consequently, in the region of the Na layer, the second term on the right hand side of (41) is negligible so that (39) can be written as

$$r_s(z,t) \simeq \frac{-1}{(\gamma-1)}\left[1 - \frac{\gamma H(z-z_o)}{\sigma_o^2}\right] r_a(z,t)\,. \tag{42}$$

By using (42) the mean-square Na density perturbations can be expressed in terms of the mean-square atmospheric density fluctuations.

$$<r_s^2(z,t)> = \left[1 - \gamma H(z-z_o)/\sigma_o{}^2\right]^2 <r_a^2(z,t)> / \left(\gamma-1\right)^2\,. \tag{43}$$

The vertical wavenumber and temporal frequency spectra of the atmospheric and Na density perturbations can also be related using (42). For the temporal spectrum we have

$$\frac{<|R_a(z,\omega)|^2>}{T} = F_a(\omega) \tag{44}$$

$$\frac{<|R_s(z,\omega)|^2>}{T} = \frac{\left[1-\gamma H(z-z_o)/\sigma_o{}^2\right]^2}{(\gamma-1)^2}\; F_a(\omega), \tag{45}$$

where R_a and R_s are respectively the temporal Fourier transforms of r_a and r_s.

$$R_a(z,\omega) = \int_o^T r_a(z,t)\, e^{i\omega t}\, dt \tag{46}$$

$$R_s(z,\omega) = \int_o^T r_s(z,t)\, e^{i\omega t}\, dt \tag{47}$$

$F_a(\omega)$ is the atmospheric density perturbation spectrum and T is the observation period. If the vertical extent of the Na layer L is large compared to the vertical correlation length of the density or wind perturbations, then the vertical wavenumber spectra of the Na and atmospheric density fluctuations are also related,

$$\frac{<|R_a(k_z,t)|^2>}{L} = F_a(k_z) \tag{48}$$

$$\frac{<|R_s(k_z,t)|^2>}{L} \simeq \frac{1}{\mu^2}\, F_a(k_z) \tag{49}$$

where

$$\mu^2 = \frac{(\gamma-1)^2}{\left[1+\left(\gamma HL/\sigma_o{}^2\right)^2/12\right]} \tag{50}$$

and

$$R_a(k_z,t) = \int_{z_0-L/2}^{z_0+L/2} r_a(z,t)\, e^{ik_z z}\, dz \tag{51}$$

$$R_s(k_z,t) = \int_{z_0-L/2}^{z_0+L/2} r_s(z,t)\, e^{ik_z z}\, dz\ . \tag{52}$$

Although Rayleigh and Na lidars measure density, the gravity wave polarization and dispersion relations can be used to relate the mean-square atmospheric density perturbations $\left(<r_a^2> = <(\Delta\rho_a/\rho_a)^2>\right)$ and density power spectra (F_a) to the mean-square horizontal wind perturbations $\left(<u^2>\right)$ and horizontal wind power spectra (F_u).

$$<u^2(z,t)> \simeq (g/N)^2 <r_a^2(z,t)> = (g/N)^2 < \left[\frac{\Delta\rho_a(z,t)}{\rho_a(z,t)}\right]^2 > \tag{53}$$

$$F_u(k_z) \simeq (g/N)^2\, F_a(k_z) \tag{54}$$

$$F_u(\omega) \simeq (g/N)^2\ F_a(\omega) \tag{55}$$

V. ANALYSIS OF LIDAR DENSITY DATA FOR DYNAMICS STUDIES

In this section, we discuss the problem of estimating the mean-square atmospheric density fluctuations and spectra from the Rayleigh and Na lidar data. Because the parameters r_a and r_s are derived from photon count profiles, they will be contaminated by shot noise. The statistical fluctuations of the photon counts establish noise floors in the power spectra and bias terms in the computed mean-square density variations which are inversely proportional to the signal count. Photon noise will also introduce random errors into the measured values of the spectra and mean-square density variations. The variances of these parameters will also be inversely related to the signal count. In many cases the noise floors, bias terms and variances can be reduced substantially by spatially and temporally filtering the photocount data.

The temporal variations of the Rayleigh photon count N_R caused by wave affects are typically very small. In the following analysis we assume for simplicity that N_R is constant with respect to time. If photon noise effects are included, the mean-square value of r_a is given by

$$<r_a^2(z,t)> = < \left[\frac{\Delta\rho_a(z,t)}{\rho_a(z,t)}\right]^2 > + \frac{1}{N_R(z)}\ . \tag{56}$$

The first term on the right-hand side of (56) is the mean-square atmospheric density fluctuations and the second term is the contribution due to photon noise. Recall the $N_R(z)$ is the photon count per profile in the range bin corresponding to altitude z. This photon noise bias term can be reduced substantially by spatially and temporally filtering r_a. Low-pass filtering increases the effective photon count because the process is equivalent to summing the counts over several adjacent range bins and profiles. Assume the filters are ideal low-pass filters with spatial and temporal cutoff frequencies given by

$$k_c = 2\pi/\lambda_c \leq \pi/\Delta z \tag{57}$$

$$\omega_c = 2\pi/T_c \leq \pi/\Delta t \tag{58}$$

where λ_c is the cutoff wavelength and T_c is the cutoff period. The upper limits on the cutoff frequencies given in (57) and (58) are established by the Nyquist sampling criterion. If the bandwidths of these filters are larger than the bandwidths of the atmospheric density fluctuations, then after filtering the mean-square value of r_a is

$$<r_a^2(z,t)> \simeq < \left[\frac{\Delta\rho_a(z,t)}{\rho_a(z,t)}\right]^2 > + \frac{4\Delta z\Delta t}{\lambda_c T_c N_R(z)} . \tag{59}$$

In processing the lidar data, ensemble averages must be approximated by computing either spatial or temporal averages or a combination of both. The mean-square density perturbations as a function of altitude can be estimated by averaging r_a^2 over the observation period. The expected value (<•>) of the temporal average is

$$\frac{1}{T}\int_0^T <[r_a(z,t) - \bar{r}_a(z)]^2> dt = < \left[\frac{\Delta\rho_a(z)}{\rho_a(z)}\right]^2 > + \frac{4\Delta z\Delta t}{\lambda_c T_c N_R(z)} \tag{60}$$

$$\bar{r}_a(z) = \frac{1}{T}\int_0^T r_a(z,t)dt \tag{61}$$

The value of the density perturbations calculated using (60) includes the affects of all waves with periods between T_c and T and vertical wavelengths greater than λ_c. The effects of waves with vertical wavelengths greater than λ_o can be eliminated if $\bar{r}_a$ and $(r_a-\bar{r}_a)^2$ are also averaged over the height range $(z-\lambda_o/2,\ z+\lambda_o/2)$. In this case the calculated density perturbations will include the effects of all waves with periods between T_c and T and vertical wavelengths between λ_c and λ_o.

Similarly, the temporal variations of the mean-square density perturbations can be estimated by averaging r_a^2 over the altitude range of observations. The expected value (<•>) of the spatial average is

$$\frac{1}{L}\int_{z_s-L/2}^{z_s+L/2} <[r_a(z,t) - \bar{r}_a(t)]^2> dz = < \left[\frac{\Delta\rho_a(t)}{\rho_a(t)}\right]^2 > + \frac{4\Delta z\Delta t}{\lambda_c T_c}\frac{1}{L}\int_{z_s-L/2}^{z_s+L/2}\frac{dz}{N_R(z)} \tag{62}$$

$$\bar{r}_a(t) = \frac{1}{L}\int_{z_s-L/2}^{z_s+L/2} r_a(z,t)\,dz\ . \tag{63}$$

The density perturbations calculated using (62) include the effects of all waves with periods greater than T_c and vertical wavelengths between λ_c and L. If $\bar{r}_a$ and $(r_a-\bar{r}_a)^2$ are also averaged over the period $(t-T_o/2,\ t+T_o/2)$, the calculated value of mean-square density perturbations will include the affects of all waves with periods between T_c and T_o and vertical wavelengths between λ_c and L.

The corresponding results for the mean-square atmospheric density fluctuations, calculated from the Na density data are

$$\frac{(\gamma-1)^2}{[1-\gamma H(z-z_o)/\sigma_o^2]^2}\frac{1}{T}\int_0^T < \left[r_s(z,t) - \bar{r}_s(z)\right]^2 > dt =$$

$$< \left[\frac{\Delta\rho_a(z)}{\rho_a(z)}\right]^2 > + \frac{(\gamma-1)^2}{\left[1-\gamma H(z-z_o)/\sigma_o^2\right]^2} \frac{4\Delta z\Delta t}{\lambda_c T_c N_s(z)} \tag{64}$$

$$\frac{\mu^2}{L} \int_{z_o-L/2}^{z_o+L/2} < [r_s(z,t) - \bar{r}_s(t)]^2 > dz = < \left[\frac{\Delta\rho_a(t)}{\rho_a(t)}\right]^2 >$$

$$+ \frac{4\Delta z\Delta t}{\lambda_c T_c} \frac{\mu^2}{L} \int_{z_o-L/2}^{z_o+L/2} \frac{dz}{N_s(z)} \tag{65}$$

where

$$\bar{r}_s(z) = \frac{1}{T} \int_0^T r_s(z,t)\, dt \tag{66}$$

$$\bar{r}_s(t) = \frac{1}{L} \int_{z_o-L/2}^{z_o+L/2} r_s(z,t)\, dt\ . \tag{67}$$

Notice that at the altitude $z_1 = z_o + \sigma_o^2/\gamma H$, the multiplicative factor on the left-hand side of (64) becomes infinite. This is the altitude at which first-order wave effects (i.e., effects related to $\underline{V}$) have no influence on the Na profile [Gardner and Shelton, 1985]. In the vicinity of this altitude, the calculations in (64) are very sensitive to noise so that the atmospheric density fluctuations cannot be estimated accurately. Because σ_o is nominally 4.25 km and H is approximately 6 km at the mesopause, z_1 is typically 2.4 km above the layer centroid height or about 95 km.

The appropriate cutoff frequencies for filtering r_a or r_s can be determined by computing the power spectra of the atmospheric density fluctuations. The statistical fluctuations of the photon counts establish noise floors in the spectra which are inversely proportional to the signal count. The noise floors in the temporal power spectra can be reduced by first spatially low-pass filtering r_a and r_s. If photon noise effects are included and the temporal spectra of r_a are computed at each altitude in the Rayleigh region according to (44) and (46) and then averaged, we have for the expected value (<•>) of the spatial average

$$\frac{1}{L} \int_{z_s-L/2}^{z_s+L/2} \frac{<|R_a(z,\omega)|^2>}{T} dz = F_a(\omega) + \frac{2\Delta z\Delta t}{\lambda_c} \frac{1}{L} \int_{z_s-L/2}^{z_s+L/2} \frac{dz}{N_R(z)}\ . \tag{68}$$

Similarly, if the spectra of r_s are computed throughout the Na layer according to (45) and (47) and then averaged, we obtain

$$\frac{\mu^2}{L} \int_{z_o-L/2}^{z_o+L/2} \frac{<|R_s(z,\omega)|^2>}{T} dz \simeq F_a(\omega) + \frac{2\Delta z\Delta t}{\lambda_c} \frac{\mu^2}{L} \int_{z_o-L/2}^{z_o+L/2} \frac{dz}{N_s(z)}\ . \tag{69}$$

Equations (68) and (69) were derived by assuming that r_a and r_s were low-pass filtered spatially with a cut-off wavelength λ_c before the spectra were computed. The corresponding results for the vertical wave number spectra computed from each profile according to (48)-(52) and then averaged over the observation period are

$$\frac{1}{T}\int_0^T \frac{<|R_a(k_z,t)|^2>}{L}\,dt \simeq F_a(k_z) + \frac{2\Delta z\Delta t}{T_c}\,\frac{1}{L}\int_{z_S-L/2}^{z_s+L/2}\frac{dz}{N_R(z)} \tag{70}$$

$$\frac{\mu^2}{T}\int_0^T \frac{<|R_s(k_z,t)|^2>}{L}\,dt \simeq F_a(k_z) + \frac{2\Delta z\Delta t}{T_c}\,\frac{\mu^2}{L}\int_{z_S-L/2}^{z_s+L/2}\frac{dz}{N_s(z)}\,. \tag{71}$$

Equations (70) and (71) were derived by assuming that r_a and r_s were low-pass filtered temporally with a cut-off period T_c before the spectra were computed.

The bias terms and noise floors all depend inversely on the count levels. If we assume that the atmospheric density profile is approximately exponential and the Na density profile is approximately Gaussian, then

$$N_R(z) \simeq N_R(z_s)\,(z_s/z)^2\,e^{-(z-z_s)/H} \tag{72}$$

$$N_s(z) \simeq \frac{N_s\Delta z}{\sqrt{2\pi}\sigma_o}\,e^{-(z-z_o)^2/2\sigma_o^2} \tag{73}$$

and

$$\frac{1}{L}\int_{z_S-L/2}^{z_s+L/2}\frac{dz}{N_R(z)} \simeq \frac{H}{LN_R(z_s)}\left(1+\frac{L/2-H}{z_s}\right)^2 e^{L/2H} \tag{74}$$

$$\frac{1}{L}\int_{z_O-L/2}^{z_o+L/2}\frac{dz}{N_s(z)} \simeq \frac{\sqrt{2\pi}\sigma_o}{N_s\Delta z}\sum_{n=0}^{\infty}\left(\frac{L}{2\sqrt{2}\sigma_o}\right)^{2n}\Big/(2n+1)n!$$

$$\simeq \frac{4\sqrt{2\pi}\sigma_o}{N_s\,\Delta z}\,. \tag{75}$$

In (75) we have assumed that L is the altitude range over which the Na density is greater than 5 percent of the peak density. In this case, L is approximately 20 km. N_s and $N_R(z_s)$ are evaluated for typical system parameters in equations (6) and (14), respectively.

The statistical fluctuations of the photon counts will also contribute random errors to the measured values of the density perturbations and spectra. The variances of $<(\Delta\rho_a/\rho_a)^2>$ and F_a caused by photon noise are similar in form to the bias terms and photon noise floors and are inversely proportional to the signal counts. The biases and variances of the computed mean-square atmospheric density fluctuations are summarized in Table I. The noise floors and variances of the computed atmospheric density spectra are summarized in Table II. In all cases, the bias terms, noise floors and variances are inversely proportional to the lidar performance factor P_LA_R. Notice

Table I
Summary of Bias Terms and Variances of the Computed Values of the Mean-Square Relative Atmospheric Density Fluctuations

Lidar Type	Mean-Square Relative Density	Bias Term*	Mean-Square Relative Density Variance
Rayleigh	$<\left[\frac{\Delta\rho_s(z)}{\rho_s(z)}\right]^2>$	$\frac{4\Delta z\Delta t}{\lambda_c T_c N_R(z)} - \frac{70}{P_L A_R \lambda_c T_c}(z/z_s)^2 e^{(z-z_s)/H}$	$<\left[\frac{\Delta\rho_s(z)}{\rho_s(z)}\right]^2> \frac{8\Delta z\Delta t}{\lambda_c T\, N_R(z)}$
Rayleigh	$<\left[\frac{\Delta\rho_s(l)}{\rho_s(l)}\right]^2>$	$\frac{4\Delta z\Delta t}{\lambda_c T_c}\frac{1}{L}\int_{z_s-L/2}^{z_s+L/2}\frac{dz}{N_R(z)} - \frac{200}{P_L A_R \lambda_c T_c}$	$<\left[\frac{\Delta\rho_s(l)}{\rho_s(l)}\right]^2> \frac{8\Delta z\Delta t}{L\ T_c}\frac{1}{L}\int_{z_s-L/2}^{z_s+L/2}\frac{dz}{N_R(z)}$
Na	$<\left[\frac{\Delta\rho_s(z)}{\rho_s(z)}\right]^2>$	$\frac{4\Delta z\Delta t}{\lambda_c T_c N_s(z)}\frac{(\gamma-1)^2}{\left[1-\gamma H(z-z_o)/\sigma_o{}^2\right]^2} - \frac{0.2}{P_L A_R \lambda_c T_c}\frac{e^{(z-z_o)^2/2\sigma_o 2}}{\left[1-\gamma H(z-z_o)/\sigma_o{}^2\right]^2}$	$<\left[\frac{\Delta\rho_s(z)}{\rho_s(z)}\right]^2> \frac{8\Delta z\Delta t}{\lambda_c T\, N_s(z)}\frac{(\gamma-1)^2}{\left[1-\gamma H(z-z_o)/\sigma_o{}^2\right]^2}$
Na	$<\left[\frac{\Delta\rho_s(l)}{\rho_s(l)}\right]^2>$	$\frac{4\Delta z\Delta t}{\lambda_c T_c}\frac{\mu^2}{L}\int_{z_o-L/2}^{z_o+L/2}\frac{dz}{N_s(z)} - \frac{0.1}{P_L A_R \lambda_c T_c}$	$<\left[\frac{\Delta\rho_s(l)}{\rho_s(l)}\right]^2> \frac{8\Delta z\Delta t}{L\ T_c}\frac{\mu^2}{L}\int_{z_o-L/2}^{z_o+L/2}\frac{dz}{N_s(z)}$

* For Rayleigh lidar data, bias terms were calculated assuming $z_s = 45$ km, $L = 30$ km and $H = 7$ km.
For Na lidar data, bias terms were calculated assuming $\sigma_o = 4.25$ km, $L = 20$ km and $H = 6$ km.
In both cases, it was assumed $\eta = 5\%$, $T_a = 0.7$ and $\lambda_L = 589$ nm.
For the bias terms the parameter units are $P_L(w)$, $A_R(m^2)$, $\lambda_c(m)$ and $T_c(s)$.

Table II
Summary of Noise Floors and Variances of the Computed Spectra of the Relative Atmospheric Density Fluctuations

Lidar Type	Relative Density Spectrum	Noise Floor*	Spectrum Variance
Rayleigh	$F_a(k_z)$	$\frac{2\Delta z\Delta t}{T_c}\frac{1}{L}\int_{z_s-L/2}^{z_s+L/2}\frac{dz}{N_R(z)} \sim \frac{100}{P_L A_R T_c}$ (m)	$F_a(k_z)\frac{2\Delta z\Delta t}{T}\frac{1}{L}\int_{z_s-L/2}^{z_s+L/2}\frac{dz}{N_R(z)}$
Rayleigh	$F_a(\omega)$	$\frac{2\Delta z\Delta t}{\lambda_c}\frac{1}{L}\int_{z_s-L/2}^{z_s+L/2}\frac{dz}{N_R(z)} \sim \frac{100}{P_L A_R \lambda_c}$ (s)	$F_a(\omega)\frac{2\Delta z\Delta t}{L}\frac{1}{L}\int_{z_s-L/2}^{z_s+L/2}\frac{dz}{N_R(z)}$
Na	$F_a(k_z)$	$\frac{2\Delta z\Delta t}{T_c}\frac{\mu^2}{L}\int_{z_0-L/2}^{z_0+L/2}\frac{dz}{N_s(z)} \sim \frac{0.05}{P_L A_R T_c}$ (m)	$F_a(k_z)\frac{2\Delta z\Delta t}{T}\frac{\mu^2}{L}\int_{z_0-L/2}^{z_0+L/2}\frac{dz}{N_s(z)}$
Na	$F_a(\omega)$	$\frac{2\Delta z\Delta t}{\lambda_c}\frac{\mu^2}{L}\int_{z_0-L/2}^{z_0+L/2}\frac{dz}{N_s(z)} \sim \frac{0.05}{P_L A_R \lambda_c}$ (s)	$F_a(\omega)\frac{2\Delta z\Delta t}{L}\frac{\mu^2}{L}\int_{z_0-L/2}^{z_0+L/2}\frac{dz}{N_s(z)}$

* Noise floors were calculated using the same parameter values listed in Table I.
For the noise floors the parameter units are P_L(w), A_R(m^2), λ_c(m) and T_c(s).

that for comparable performance factors, these parameters for the Na data are approximately 3 orders of magnitude smaller than the corresponding values for the Rayleigh data. This performance advantage for Na lidars is a consequence of the large resonant backscatter cross-section for Na and the large density gradients on the bottom- and top-sides of the layer which enhance the wind induced density perturbations. Na lidars have an additional advantage because the atmospheric density fluctuations are usually much larger near the mesopause compared to the stratopause. However, the most powerful Na lasers available today can only achieve power levels ~5 w compared to 50 w for the near-UV excimer lasers which are beginning to be used for larger Rayleigh lidars. Ths most powerful Na lidars currently in operation have performance factors in the range of 5-10 wm^2 compared to almost 100 wm^2 for the most powerful Rayleigh lidars. From a scientific standpoint, performance comparisons between the two techniques are largely irrelevant since Rayleigh and Na lidars make observations in different but equally important atmospheric regions.

Because the relative atmospheric density fluctuations range from at most a few percent in the upper stratosphere to 5-15 percent near the mesopause, the bias terms should be smaller than 10^{-4} to insure accurate estimates of the mean-square density fluctuations. This condition requires

$$P_L A_R \lambda_c T_c > 10^6 \text{ wm}^3\text{s} \qquad \text{(Rayleigh lidars)} \tag{76}$$

$$P_L A_R \lambda_c T_c > 10^3 \text{ wm}^3\text{s} \qquad \text{(Na lidars)}\,. \tag{77}$$

For a Rayleigh lidar with $P_L = 10$ w and $A_R = 1$ m^2, this condition can be met by low-pass filtering the data using a cutoff period of 100 s and a cutoff wavelength of 1 km. For a Na lidar with $P_L = 1$ w and $A_R = 1$ m^2, this condition can be met by filtering using a 10 s cutoff period and a 100 m cutoff wavelength.

In recent years, experimental data and theoretical arguments have suggested that there is a universal form for the vertical wavenumber spectrum of horizontal wind perturbations in the upper atmosphere [Dewan and Good, 1986; Smith et al., 1989]. The theoretical horizontal wind spectrum is

$$F_u(k_z) \sim \frac{N^2}{2k_z^3}\,. \tag{78}$$

The theoretical spectrum for the atmospheric density fluctuations can be derived using (54) and (78)

$$F_a(k_z) \sim \left(\frac{\gamma\text{-}1}{\gamma H}\right)^2 \frac{1}{2k_z^3}\,. \tag{79}$$

At a vertical wavelength of 1 km, (79) gives a value of approximately 5×10^{-3} m for the density spectrum. For a lidar to observe reliably the spectrum out to a spatial frequency of 1 km^{-1}, the noise floors must be less than 5×10^{-3} m which requires (see Table II)

$$P_L A_R T_c > 2\times10^4 \text{ wm}^2\text{s} \qquad \text{(Rayleigh lidars)} \tag{80}$$

$$P_L A_R T_c > 10 \text{ wm}^2\text{s} \qquad \text{(Na lidars)}\,. \tag{81}$$

This criterion can be satisfied by a Rayleigh lidar with $P_L = 10$ w, $A_R = 1$ m^2 and $T_c = 30$ min or a Na lidar with $P_L = 1$ w, $A_R = 1$ m^2 and $T_c = 10$ s.

No theoretical models have yet been developed for the gravity wave temporal spectrum. However, radar and lidar observations in the mesosphere indicate that at the buoyancy frequency (~ 0.2 min^{-1}), the magnitude of the wave spectrum is typically of the order of 10^3 - 10^4 m^{2}s^{-1}. According to (55), this level corresponds to a density spectrum value of 5×10^{-3} - 5×10^{-2} s. Thus, to observe reliably the temporal spectrum out to the buoyancy frequency, the noise floors must be less than 5×10^{-3} s which requires (see Table II)

$$P_L A_R \lambda_c > 2\times10^4 \text{ wm}^3 \qquad \text{(Rayleigh lidars)} \qquad (82)$$

$$P_L A_R \lambda_c > 10 \text{ wm}^3 \qquad \text{(Na lidars)}\,. \qquad (83)$$

This criterion can be met by a Raleigh lidar with $P_L = 10$ w, $A_R = 1$ m^2 and $\lambda_c = 2$ km or a Na lidar with $P_L = 1$ w, $A_R = 1$ m^2 and $\lambda_c = 10$ m.

VI. ANALYSIS OF LIDAR DENSITY DATA FOR TEMPERATURE STUDIES

Both Rayleigh and Na lidars can be used to measure atmospheric temperature profiles. The Rayleigh technique was pioneered by Elterman [1951, 1953 and 1954] in the early 1950s. The temperature profile is inferred from the relative atmospheric density profile by using the hydrostatic equation and the ideal gas law. The Na technique was first demonstrated in the late 1970s by Gibson et al. [1979]. Na temperature is inferred by observing the thermally broadened hyperfine structure of the D_2 emission line with a narrowband laser. In this section, we discuss the problem of estimating the atmospheric temperature profile from the Rayleigh and Na lidar data. The impact of signal photon noise on the accuracies of the inferred temperatures is evaluated.

Rayleigh lidars measure the relative atmospheric density profile. To determine temperature from these measurements, the density profile is integrated downward using the hydrostatic equation, ideal gas law and an assumed upper level temperature. The initial upper level temperature may be chosen from a model because when the equation has been integrated downward by about one and one-half scale heights, the calculated temperature is relatively insensitive to the initial estimate. The hydrostatic equation

$$dP = -\rho_a\, g\, dz \qquad (84)$$

can be combined with the ideal gas law

$$P = \rho_a RT/M \qquad (85)$$

and integrated to yield

$$T(z) = \frac{T(z_1)\,\rho_a(z_1)}{\rho_a(z)} + \frac{M}{R}\int_z^{z_1} \frac{g(r)\rho_a(r)}{\rho_a(z)}\,dr \qquad (86)$$

where

$T(z)$ = atmospheric temperature profile (K)
$P(z)$ = atmospheric pressure profile (mbar)
$\rho_a(z)$ = atmospheric density profile (m^{-3})
z_1 = altitude of upper level temperature estimate (m).

Notice that the atmospheric density appears as a ratio of densities in (86). Thus, to determine temperature it is only necessary to measure the relative density. The accuracy of this technique depends upon the molecular or Rayleigh photocount $N_R(z)$ and the accuracy of the upper level temperature estimate $T(z_1)$. Accuracy can be improved substantially by low-pass filtering the photocount data. When filtering the data spatially, care must be taken to insure that the filtering process does not distort the atmospheric density profile. The most effective approach is to filter

$r_a(z,t)$ rather than the photocount profiles. In this case the relative atmospheric density profile is estimated from the filtered values of r_a using (25). If r_a is filtered using ideal low-pass spatial and temporal filters, then the variance of the inferred temperature profile is

$$\mathrm{var}[T(z)] \simeq \frac{4\Delta z\Delta t}{\lambda_c T_c}\,\frac{T^2(z)}{N_R(z)} + \left\{\mathrm{var}[T(z_1)] + \frac{4\Delta z\Delta t}{\lambda_c T_c}\,\frac{T^2(z_1)}{N_R(z_1)}\right\}\exp[-2(z_1-z)/H]. \qquad (87)$$

Notice that the error is inversely proportional to the photon count. For altitudes well below z_1, the standard deviation of the temperature measurement is

$$\mathrm{Std}[T(z)] \simeq \left[\frac{4\Delta z\Delta t}{\lambda_c T_c N_R(z)}\right]^{1/2} T(z) \sim \frac{8T(z)}{\sqrt{P_L A_R \lambda_c T_c}}\,(z/z_s)e^{(z-z_s)/2H}. \qquad (88)$$

Near the stratopause the temperature is approximately 270 K. To achieve a measurement accuracy of 1 K at $z=z_s$ requires

$$P_L A_R \lambda_c T_c \gtrsim 5\mathrm{x}10^6\ \mathrm{wm^3s}. \qquad (89)$$

This accuracy can be achieved by using a lidar with P_L=10 w, A_R = 1 m^2, λ_c = 1 km and T_c = 10 min.

Because the atmospheric density and hence the signal level decreases exponentially with increasing altitude, it is very difficult to extend Rayleigh lidar temperature measurements to high altitudes. For a fixed temperature accuracy, each factor of 10 improvement in laser power or telescope area increases the altitude by only about 2 scale heights. Because the molecular backscatter cross section is approximately proportional to λ^{-4} (λ is the optical wavelength), modern Rayleigh lidars use high power (10-50 W) short wavelength lasers such as frequency doubled Nd:YAG lasers operating at 532 nm or XeF excimer lasers operating at 351 nm and large telescopes (~ 1 m^2) to obtain temperature profiles up to about 90 km with integration times ranging from 5 min to 1 h depending on the altitude. The relatively lower power (1-5 W) long wavelength (589 nm) Na lidars are only capable of achieving useful Rayleigh temperature measurements up to about 70 km.

Another effective technique for obtaining temperature profiles near the mesopause is to observe the thermally broadened hyperfine structure of the Na D_2 emission line with a narrowband lidar. The Na D_2 resonance line is actually a Doppler broadened doublet composed of 6 hyperfine lines. Figure 3 is a plot of the resonance line as a function of wavelength for several values of temperature. As the temperature decreases the minimum between the D_{2a} and D_{2b} lines decreases and the peak intensities of the D_{2a} and D_{2b} lines increase. The ratio of the backscatter cross section at the minimum between the two peaks to the value at the D_{2a} peak is a very sensitive indicator of temperature. This ratio can be measured by tuning a narrowband lidar alternately between the peak and minimum and the data used to deduce the Na temperature profile [Fricke and von Zahn, 1985]. Neuber et al. [1988] have used a variation of this technique to obtain high resolution temperature profiles of the mesopause region above Andoya, Norway (69°N, 15°E).

If we neglect saturation effects and extinction of the laser beam as it propagates through the Na layer, the Na backscatter cross section can be referred from the photocount profile by using equation (5)

$$\sigma_{\mathrm{eff}} = \frac{z^2\sigma_R\rho_a(z_R)}{z_R{}^2\rho_s(z)}\,\frac{\left[N_s(z) - \hat{N}_B R_L\Delta t\right]}{\left[N_R(z_R) - \hat{N}_B R_L\Delta t\right]}. \qquad (90)$$

In practice saturation and extinction effects can be significant and must be taken into account in processing the Na temperature data [Simonich and Clemesha, 1983; Welsh and Gardner, 1989]. If we assume for simplicity that background noise is negligible, then the ratio of the effective backscatter cross section at the frequency of the minimum (f_m) between the D_{2a} and D_{2b} peaks to the value at the D_{2a} frequency (f_a) is given by

$$\frac{\sigma_{eff}(f_m)}{\sigma_{eff}(f_a)} = \frac{N_s(z,f_m)}{N_s(z,f_a)}\,\frac{N_R(z_R,f_a)}{N_R(z_R,f_m)} \simeq \frac{N_s(z,f_m)}{N_s(z,f_a)} \,. \tag{91}$$

The functional dependence of this ratio on temperature is affected by homogeneous and collisional broadening, by anomalous Zeeman splitting caused by the earth's magnetic field and by the laser line shape and line width. An excellent discussion of these effects can be found in Fricke and von Zahn [1985]. This ratio is a monotonically increasing function of temperature. It varies from approximately 0.3 for T=150 K to 0.4 for T=250 K.

The accuracy of the Na temperature measurement depends on the laser linewidth, laser tuning stability and photon noise. If the laser is tuned alternately between the D_{2a} peak and the minimum between the D_{2a} and D_{2b} peaks, the variance of the inferred Na temperature is [She et al., 1989]

$$\mathrm{var}[T(z)] \simeq \frac{4\Delta z\Delta t}{\lambda_c T_c}\,\frac{T^2(z)}{N_s(z,f_a)} \,. \tag{92}$$

In (92) Δt is the total time to acquire one profile with the laser tuned to the D_{2a} peak and one profile with the laser tuned to the minimum between the peaks. Normally it is twice the time required to collect a single profile.

Just as with the Rayleigh technique, the Na temperature error is inversely proportional to the photon count. If we substitute (6) and (73) into (92) and assume Δt in (92) is twice the value specified in (6), then the standard deviation of the temperature error is

$$\mathrm{std}[T(z)] \simeq \left[\frac{4\Delta z\Delta t}{\lambda_c T_c N_s(z)}\right]^{1/2} T(z) \sim \frac{2T(z)}{\sqrt{P_L A_R \lambda_c T_c}}\, e^{(z-z_0)^2/4\sigma_0^2} \,. \tag{93}$$

Near the mesopause the seasonal variation in temperature can be substantial. The annual mean is approximately 200 K. To achieve a measurement accuracy of 1 K at the peak of the Na layer when T = 200 K requires

$$P_L A_R \lambda_c T_c \gtrsim 2\mathrm{x}10^5\ \mathrm{wm^3 s} \,. \tag{94}$$

This accuracy can be achieved by using a lidar with P_L=1 w, A_R=1 m^2, λ_c = 1 km and T_c = 3 min. By comparison a Rayleigh lidar could achieve this same accuracy and resolution at 92 km only if $P_L A_R \sim 2 \times 10^4$ wm^2. The Rayleigh lidar would require a 1000 w laser and 5 m diameter telescope to achieve the same performance as the Na lidar!

VII. PRACTICAL CONSIDERATIONS IN PROCESSING RAYLEIGH AND NA LIDAR DATA

The quality of the lidar data is influenced by many factors including background noise and variations in the laser power, laser tuning and atmospheric transmittance which can cause substantial variations in the detected signal level. These effects must be compensated so that the true variations in atmospheric density and temperature can be measured. The most common approach for compensating variations in the signal level is to normalize the photocount profiles by the photon count at an altitude free of aerosol scattering. For this technique to be most effective, the normalizing count should be large to reduce errors caused by photon noise. In some applications, the Rayleigh count is averaged over a wide altitude range (say 30-60 km) to obtain the most stable and accurate estimate of the signal level used for normalizing the profiles.

The photocount profiles must also be accurately compensated for the effects of background noise. In well-engineered lidar systems the background count rate is constant so that the noise count per range bin can be estimated by averaging the counts in high altitude range bins where the Rayleigh and Na counts are negligible. Sometimes the large signal levels at low altitudes can cause saturation and after-pulsing in the PMT so that the count rate decreases or oscillates with time (and therefore altitude). This problem can be minimized by using PMT's designed to exhibit minimum after-pulsing and by gain switching the PMT to eliminate saturation by the strong low-altitude signals. The Rayleigh temperature algorithm (equation (86)) is especially sensitive to errors in the estimated background noise level. Reliable temperature measurements can only be obtained by taking great care to insure that the background count level is properly estimated and subtracted from the photocount profiles. Data processing for Na density studies is straightforward and has been described in detail elsewhere [Gardner and Shelton, 1981; Gardner et al., 1986].

For dynamics studies, data analysis is critically dependent on the accurate computation of the relative density variations r_a and r_s. These computations, according to equations (23) and (24), require knowledge of the unperturbed profiles or equivalently, measurements of $f_a(z)$ and $f_s(z)$. These functions can be estimated by appropriately averaging the photocount data. If we neglect background noise, equations (1), (3), (16) and (23) can be used to express the Rayleigh and Na photocount profiles in the following form

$$N(z,t) = \frac{\beta(t)}{z^2} \exp\left[-\chi - f(z-\xi)\right] \tag{95}$$

where the parameter $\beta(t)$ depends on several factors which can vary with time including laser power and atmospheric transmittance. The natural logarithm of the range-scaled photocounts is

$$\begin{aligned} \ln\left[z^2 N(z,t)\right] &= \ln\beta(t) - \chi - f(z-\xi) \\ &\simeq \ln\beta(t) - \chi - f(z) + \frac{df(z)}{dz}\,\xi \;. \end{aligned} \tag{96}$$

If equation (96) is averaged over an observation period which is long compared to the correlation times of χ and ξ_z, the result is

$$\frac{1}{T}\int_0^T \ln\left[z^2 N(z,t)\right] dt \simeq \frac{1}{T}\int_0^T \ln\beta(t)dt - f(z) \;. \tag{97}$$

The function f(z) can now be approximated by fitting an appropriate polynomial model $\hat{f}(z)$ to equation (97). Because χ and ξ are small it is just as effective (and sometimes more convenient) to average N(z,t) rather than $\ln(z^2N)$. The accuracy of the polynomial model must be significantly better than the magnitude of the actual density perturbations. For Rayleigh studies, a 3rd or 5th order polynomial in z will provide excellent fits to $f_a(z)$ depending on the observation region and period. The 3rd order fit is most effective when the altitude range and observation period are small (say < 20 km and < 4 hrs). For Na studies, a parabolic fit to $f_s(z)$ is adequate. The relative density perturbation r(z,t) is calculated by subtracting the polynomial approximation $\hat{f}(z)$ from $\ln(z^2N)$, and then computing and subtracting the spatial average to eliminate the $\ln\beta(t)$ term

$$\begin{aligned} r(z,t) &= -\chi + \frac{df(z)}{dz}\,\xi \simeq \\ &\ln\left[z^2N(z,t)\right] - \hat{f}(z) - \frac{1}{L}\int_0^L \left\{ \ln\left[z^2N(z,t)\right] - \hat{f}(z) \right\} dz \;. \end{aligned} \tag{98}$$

A similar data processing approach can be used for Rayleigh temperature studies. To improve the accuracy of the temperature measurement it is often desirable to spatially filter the photocount profiles. If the cutoff wavelength is large (i.e., greater than a few km), filtering can severely distort the density and temperature profiles. Distortion can be avoided by filtering $r_a(z,t)$ rather than the photocounts and then computing the relative atmospheric density using equation (25)

$$\rho_a(z,t) = \exp\left[\overset{*}{r}_a(z,t) - \hat{r}_a(z)\right] \tag{99}$$

where $\overset{*}{r}_a$ denotes the filtered relative density perturbations. In the following section examples of Rayleigh and Na lidar measurements are presented.

VIII. EXAMPLE RAYLEIGH AND NA LIDAR OBSERVATIONS OF ATMOSPHERIC DENSITY AND TEMPERATURE VARIATIONS

At mesopause heights the air is dense enough to incinerate most meteors and much of the meteoric dust that enter the atmosphere. Meteoric ablation is most pronounced between 70 and 120 km and is the major source of Na and other metals in the upper atmosphere. Although measurements are rare, lidar systems have occasionally observed the ablation trails of falling meteors. Two Na lidar profiles exhibiting returns from meteor trails are plotted in Fig. 4. These trails are typically about 100 m thick FWHM and are usually observed only for a few tens of seconds because the prevailing winds quickly advect the trails across the narrow laser beam [Beatty et al., 1988].

Typical gravity wave perturbations of the Na layer are illustrated in the density profiles plotted in Fig. 5. These perturbations are caused primarily by the vertical displacement of the layer by the vertical component of the gravity wave winds (see equation (21)). The wave activity illustrated in this figure is typical of that observed in the Na layer over Urbana at all times of the year. Monochromatic waves or events, as they are sometimes called, are usually evident during all or part of each night. The wavelengths, periods, and amplitudes of these monochromatic waves can be inferred directly from the Na layer profiles or from the spatial power spectra of the profiles [Gardner and Voelz, 1987].

Mesospheric winds, temperatures and densities are also perturbed randomly by the constructive and destructive interference of multiple waves generated by different sources. The random wind and density perturbations are best characterized statistically in terms of the gravity wave vertical wavenumber and temporal frequency spectra. Examples of temporal frequency (ω) and vertical wavenumber (k_z) spectra inferred from lidar measurements of Na density perturbations using equations (54), (55), (69) and (71) are plotted in Figs. 6 and 7. It is common practice to plot the single-sided wind spectra in units of $(m^2s^{-2})/(cyc/m)$ or $(m^2s^{-2})/(cyc/s)$ rather than the double-sided spectra calculated in equations (69) and (71) in units of $(m^2s^{-2})/(rad/m)$ or $(m^2s^{-2})/(rad/s)$. The double-sided spectra can be converted to single-sided spectra by multiplying by 2. The wavenumber and radian frequency spectra can be converted to spatial and temporal frequency spectra by multiplying by 2π. Thus the single-sided spatial and temporal frequency spectra plotted in Figs. 6 and 7 were obtained by multiplying equations (69) and (71) by 4π. Theoretical analyses predict a k_z^{-3} dependence for the vertical wavenumber spectrum [Dewan and Good, 1986; Smith et al., 1987]. Radar observations typically show a power law frequency dependence for the temporal spectrum with slopes ranging from -5/3 to -2 [Vincent, 1984]. Notice the enhancement of the temporal spectrum near the Brunt-Vaisala frequency. At mesopause heights, the buoyance period is approximately 5 min. The photon noise floors are clearly evident in both spectra.

Temperatures near the mesopause have been measured using a variety of techniques. Rocket launched falling spheres [Philbrick et al., 1985] and grenades [Theon et al., 1972] and airglow emissions [e.g., Sivjee and Hamwey, 1987] have all provided useful information on the temperature structure of the mesopause region. High resolution temperature profiles can also be obtained by active probing of the thermally broadened Na D_2 resonance line with a narrowband lidar. Although this technique was pioneered by Gibson et al. [1979] in the late 1970s, routine Na temperature observations were not attempted until von Zahn and his students from the University

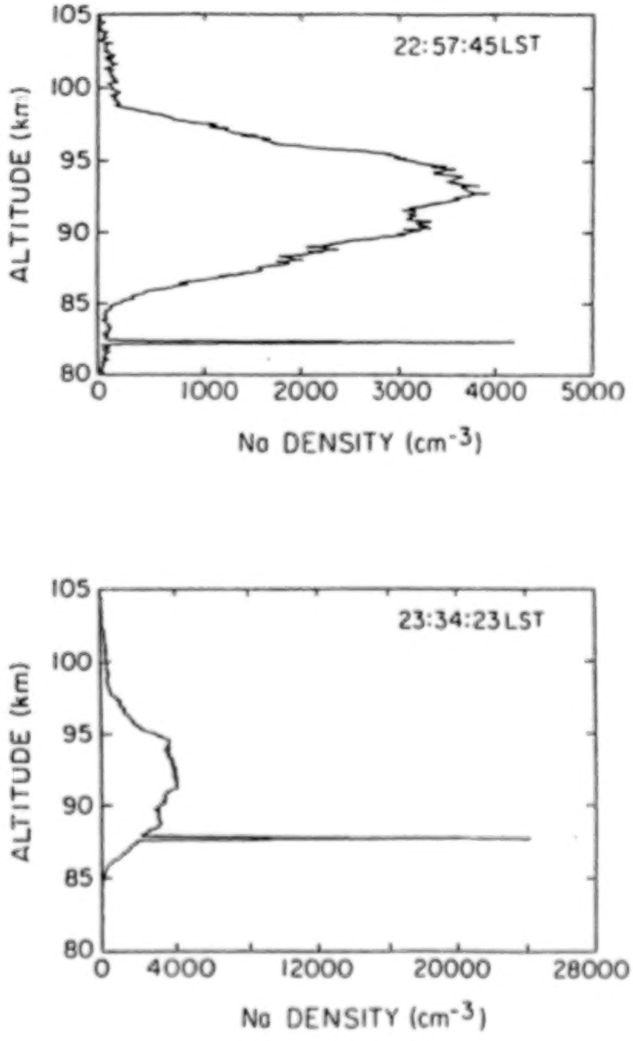


Figure 4 Na density profiles measured on March 4, 1988 at Urbana, IL. The integration period was 27s and the vertical resolution was 75 m. The dense narrow layers at 82.4 and 87 km are believed to be meteor ablation trails [Beatty et al., 1988].

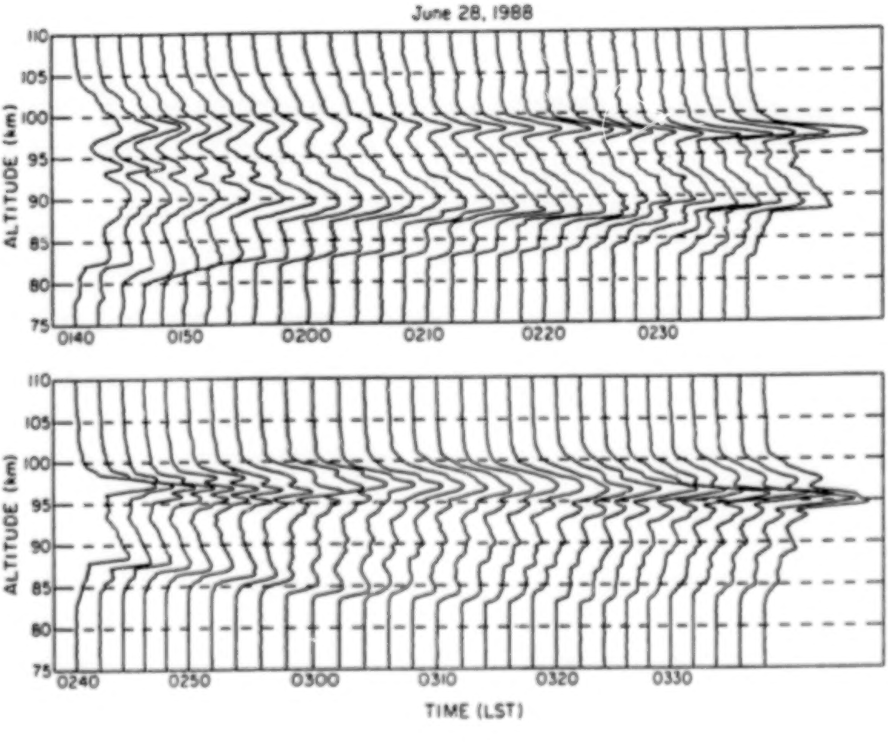


Figure 5 Na density profiles measured during the early morning of June 28, 1988 at Urbana, IL. The profiles have been plotted on a linear scale at 2 min intervals. Gravity waves are responsible for the wavelike structures in the profiles.

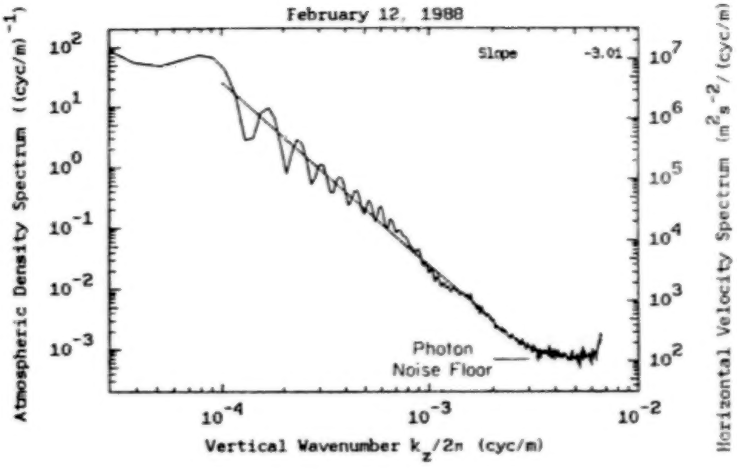


Figure 6 Vertical wavenumber spectrum of the atmospheric density and horizontal wind perturbations associated with internal gravity waves. The spectrum was inferred from Na lidar data collected between 2130 and 2330 LST on February 12, 1988 at Urbana, IL. The data were low-pass filtered temporally with a cutoff period $T_c = 5$ min before the spectrum was computed.

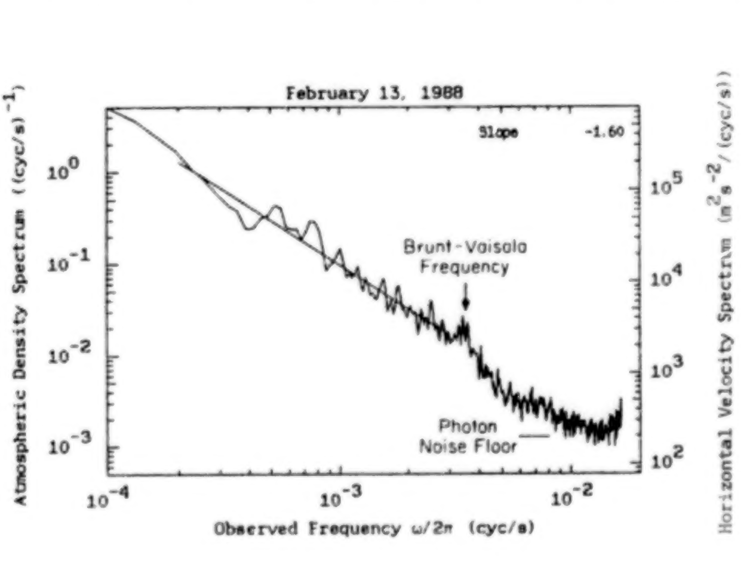


Figure 7 Temporal frequency spectrum of the atmospheric density and horizontal wind perturbations associated with internal gravity waves. The spectrum was inferred from Na lidar data collected between 0200 and 0500 LST on February 13, 1988 at Urbana,IL. The data were low-pass filtered spatially with a cutoff wavelength $\lambda_c = 1$ km before the spectrum was computed.

of Bonn improved the laser technology and began making measurements at Andoya, Norway (69°N) during the spring of 1984 [Fricke and von Zahn, 1985]. Since then, the University of Bonn group has published several excellent lidar studies of the mesopause temperature structure above Andoya [von Zahn et al., 1987; Neuber et al, 1988].

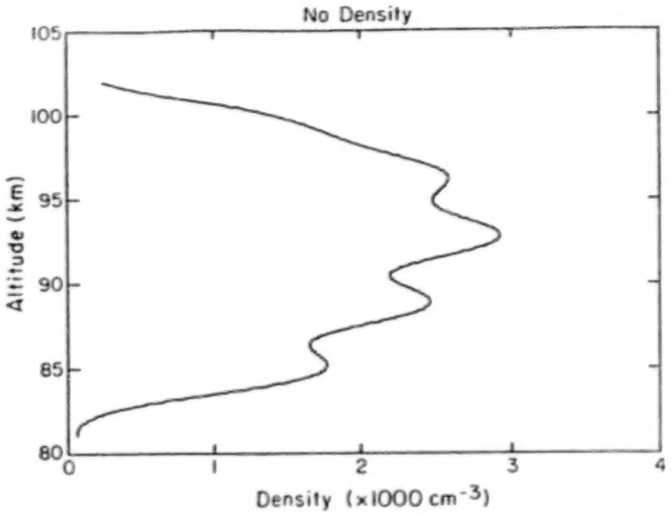
Sodium temperature and density profiles obtained at Ft. Collins, CO on the night of August 24, 1989 are plotted in Fig. 8 [She et al., 1989]. These profiles were obtained by integrating lidar returns over a period of 5 min. The mesopause near 99 km is clearly evident. The small-scale structure in the profiles is probably caused by wave dynamics. The rms measurement error near 92 km is less than 3 K. These results clearly demonstrate the impressive capabilities of Na lidar to measure the temperature structure of the mesopause region.

Significant wave activity is also frequently seen in Rayleigh lidar observations of temperature and density profiles in the lower middle atmosphere. Figure 9 is a temperature profile calculated using equation (86) and Rayleigh lidar data collected on December 10-11, 1988 at Urbana. Because the integration period is over 9 h and the vertical resolution is 5 km, the temperature error is less than 2 K throughout the observation range. The wavelike structure in the profile is caused probably by a long period gravity wave.

Figure 10 is the vertical wavenumber spectrum of density and horizontal wind perturbations for Rayleigh lidar data collected on the night of March 8-9, 1989 at Arecibo Observatory. The spectrum was calculated using equations (70) and (54) after low-pass filtering the relative density perturbations with a cutoff frequency of $(65\ \mathrm{min})^{-1}$. The altitude range extends from 30 to 50 km and the observation period extends from 2320 LST to 0545 LST. As we saw with the Na data, the photon noise floor is clearly evident in the spectrum as well as strong spectral peaks at $k_z/2\pi = (8.7\ \mathrm{km})^{-1}$ and $(5.5\ \mathrm{km})^{-1}$ which correspond to monochromatic gravity waves [Gardner et al., 1989]. The density perturbations caused by these waves are plotted versus time and altitude in Fig. 11. They were computed using the techniques described in Section VI by low-pass filtering $r_a(z,t)$. The temporal cutoff frequency was $(65\ \mathrm{min})^{-1}$. The spatial cutoff frequencies were $(2.0\ \mathrm{km})^{-1}$ for the top panel, $(3.75\ \mathrm{km})^{-1}$ for the middle panel, and $(6.6\ \mathrm{km})^{-1}$ for the bottom panel. The downward phase progression of the gravity waves is clearly evident. The relative density perturbations at 0330 LST are plotted versus altitude in Figure 12. The data were filtered temporally with a cutoff of $(65\ \mathrm{min})^{-1}$ and spatially with a cutoff of $(2\ \mathrm{km})^{-1}$. The perturbations increase from less than 1% at 30 km to almost 3% above 50 km.

The rms atmospheric density and wind perturbations at Arecibo Observatory for the nights of March 7-10 and 25-28, 1989 were calculated using (60) and (53). The average density profile for these eight nights is plotted versus altitude in Fig. 13. During March the density perturbations increased with altitude from approximately 1% ($\sim 4\ \mathrm{ms}^{-1}$) at 30 km altitude to approximately 3% ($\sim 12\ \mathrm{ms}^{-1}$) at 50 km altitude. The density and wind perturbation profiles can also be calculated from the Na data using (64) and (53). The rms atmospheric density profile calculated from Na measurements obtained at Arecibo in March 1989 is plotted in Fig. 14. This profile is the average of eight profiles measured on the same nights as the data plotted in Fig. 13. The density perturbations increased from about 2% at 80 km to about 6% near the turbopause at 100 km altitude. In fact, the decrease above 100 km suggests that for these data the turbopause may be located at 100 km altitude. The rms winds perturbations inferred from the data plotted in Figs. 13 and 14 using equations (43) and (53) are plotted in Fig. 15. The winds increase from about 4 m/s at 30 km to 25 m/s at 100 km.

The seasonal variation of gravity wave activity near the mesopause at Urbana is plotted in Fig. 16. These data were calculated by averaging the Na density perturbations throughout the layer using equations (65) and (53). The annual mean rms atmospheric density perturbation is 6.3% ($28\ \mathrm{ms}^{-1}$). There is a semi-annual oscillation in the data with an amplitude of about 1.9% ($6\ \mathrm{ms}^{-1}$). The maxima occur near the solstices and the minima near the equinoxes. Hirota [1984] also found evidence of a semi-annual oscillation of wave activity in the lower mesosphere by examining rocket measurements of temperature and density. At higher latitudes, an annual cycle was observed with a minimum in summertime. Radar measurements of gravity wave winds by Vincent and Fritts [1987] at Buckland Park (35°S) also show a semi-annual variation.



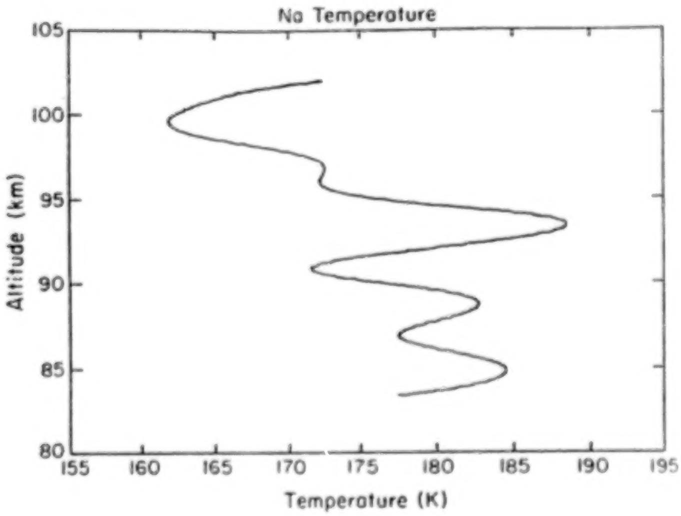


Figure 8 Na density and temperature profiles obtained at Ft. Collins, CO at 2302 LST on August 24, 1989 [She et al., 1989]. For these data $\Delta z = 1.5$ km, $\Delta T = 5$ min and the rms temperature error is less than 3 K at 92 km.

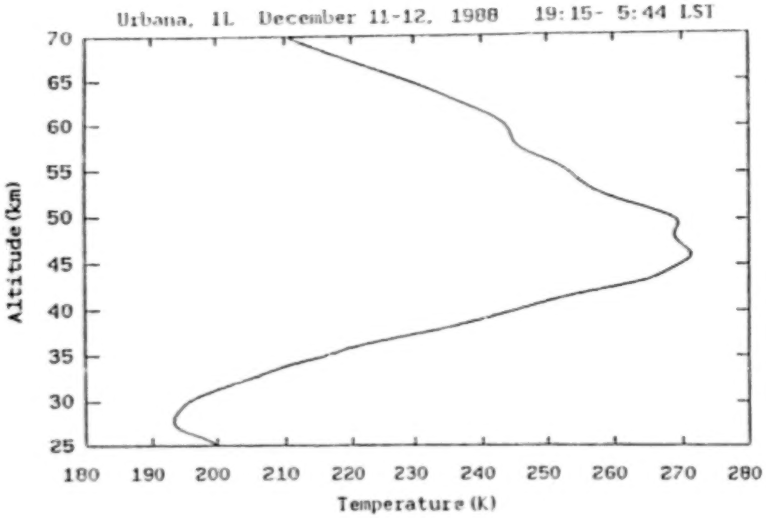


Figure 9 Temperature profile measured with a Rayleigh lidar at Urbana, IL on the night of December 10-11, 1988. For these data $\Delta z = 5$ km and $\Delta T = 9$ h 30 min. The temperature error is less than 2 K throughout the observation range.

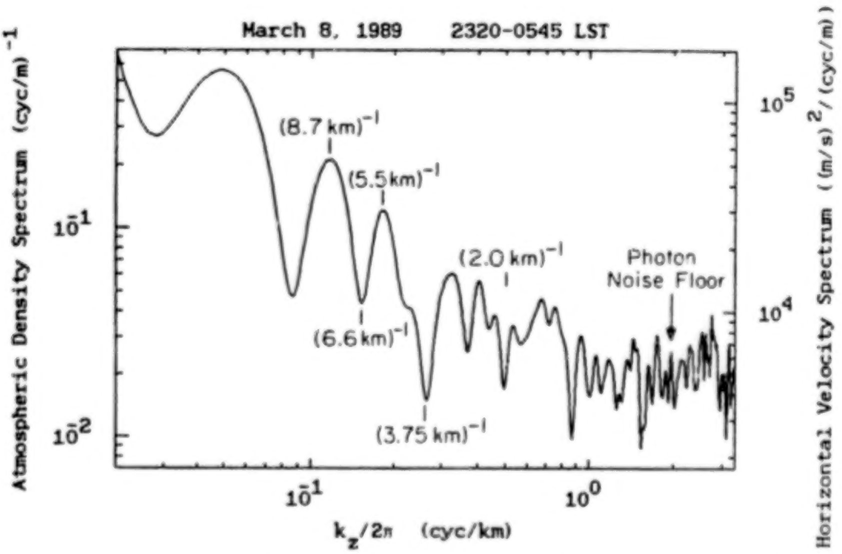


Figure 10 Vertical wavenumber spectrum of the atmospheric density and horizontal wind perturbations associated with internal gravity waves. The spectrum was inferred from Rayleigh lidar data collected between 2320 LST and 0545 LST March 8-9, 1989 at Arecibo Observatory. The data were low-pass filtered temporally with a cutoff period $T_c = 65$ min before the spectrum was computed. The altitude range is 30-50 km.

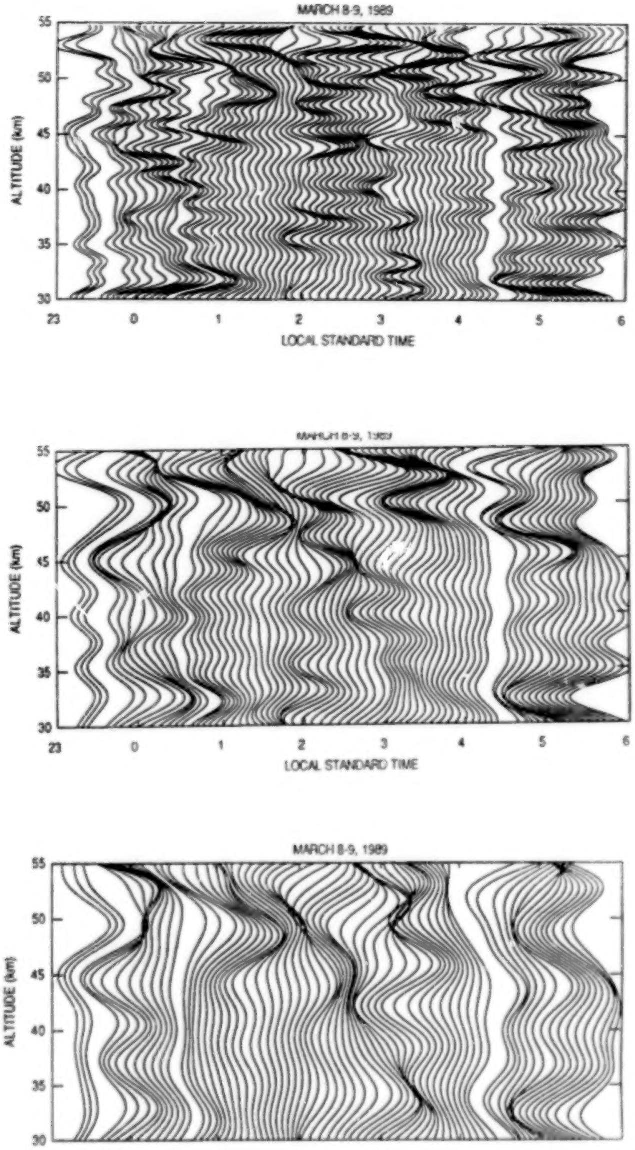


Figure 11 Altitude profiles of the relative atmospheric density perturbations measured during the night of March 8-9, 1989 at Arecibo Observatory. The data were low-pass filtered with a cutoff period $T_c = 65$ min and cutoff wavelengths $\lambda_c = 2.0$ km for the top panel, $\lambda_c = 3.75$ km for the middle panel and $\lambda_c = 6.6$ km for the bottom panel. The vertical wavenumber spectrum for these data is plotted in Fig. 10.

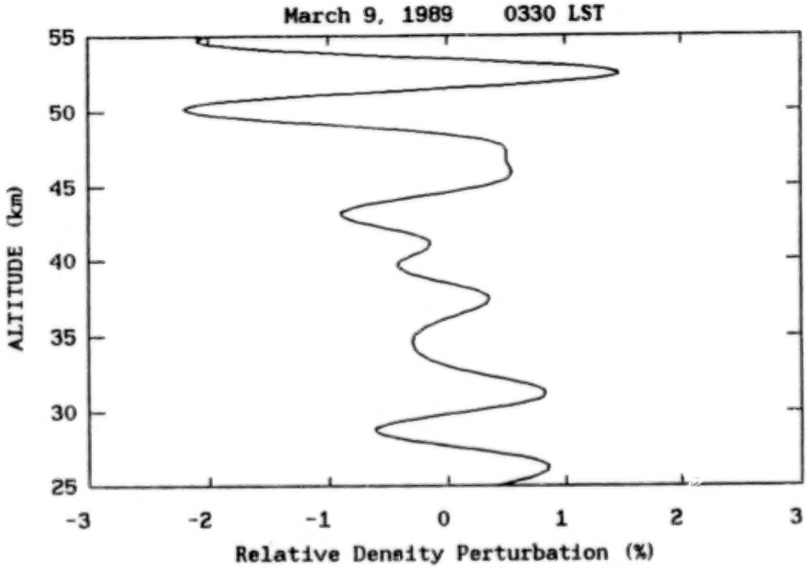


Figure 12 Altitude profile of the atmospheric relative density perturbations measured at 0330 LST on March 9, 1989 at Arecibo Observatory. The data were low-pass filtered with $\lambda_c = 2.0$ km and $T_c = 65$ min.

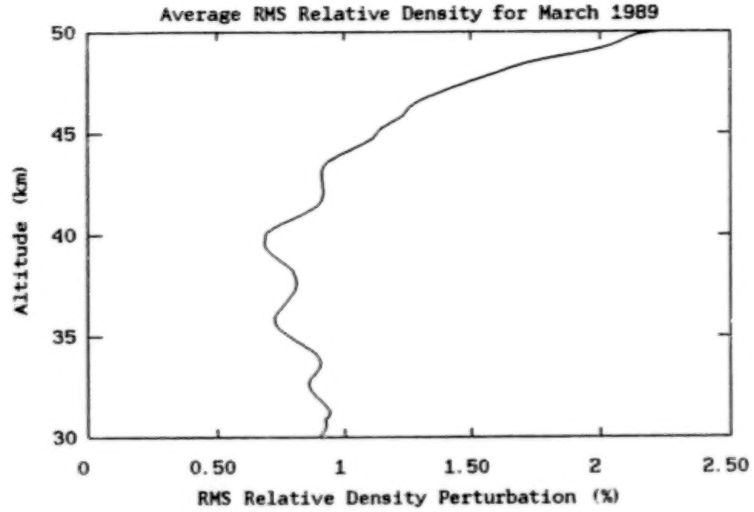


Figure 13 Altitude profile of the rms atmospheric density perturbations referred from Rayleigh lidar measurements obtained during March 1989 at Arecibo Observatory. The profile is the average of 8 nights' observations. All data were low-pass filtered with $\lambda_c = 2$ km and $T_c = 60$ min before the rms perturbations were computed.

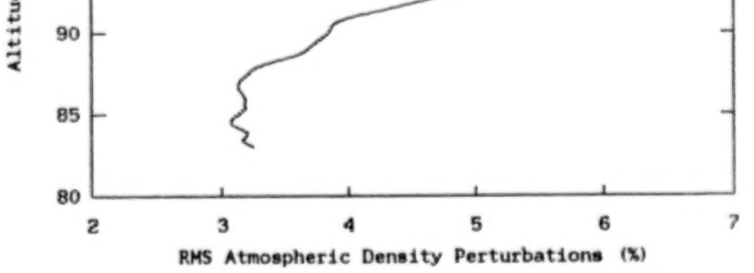


Figure 14 Altitude profile of the rms atmospheric density perturbations inferred from Na lidar measurements obtained during March 1989 at Arecibo Observatory. The profile is the average of observations obtained during the same 8 nights used to compute the data plotted in Figure 13. All data were low-pass filtered with $\lambda_c = 1$ km and $T_c = 5$ min before the rms perturbations were computed.

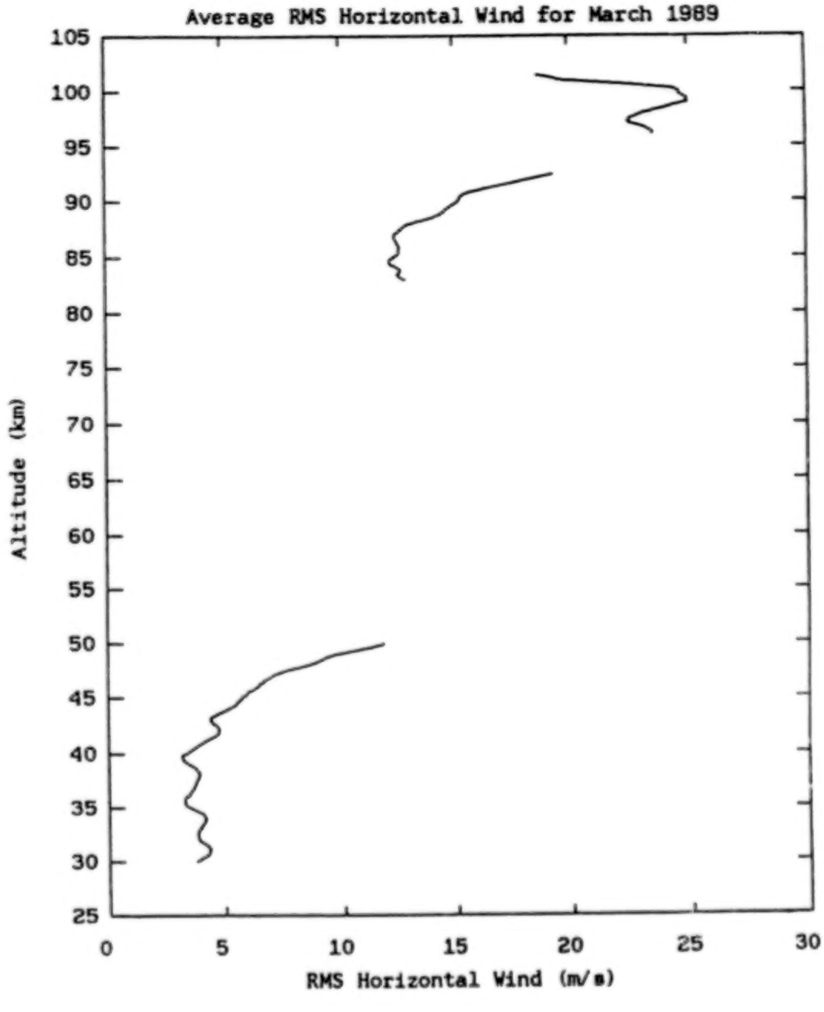
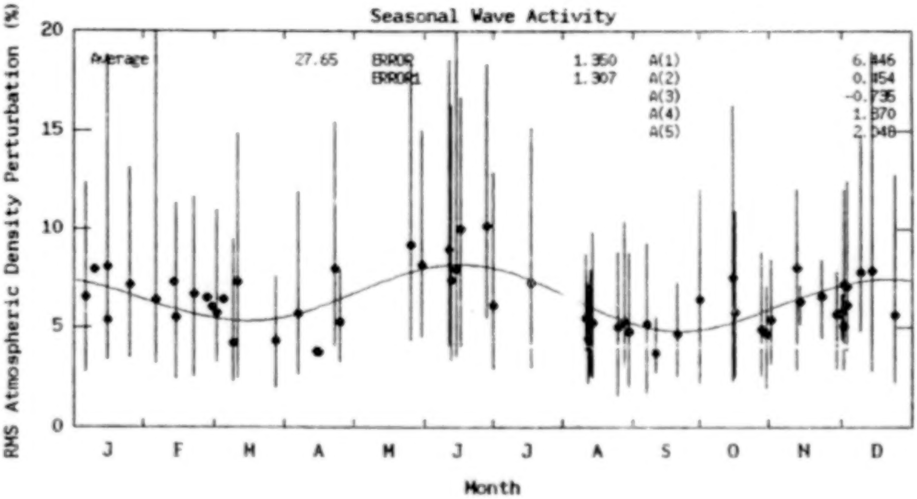


Figure 15 Altitude profile of the rms horizontal wind perturbations inferred from Rayleigh and Na lidar measurements obtained during March 1989 at Arecibo Observatory. The profile was inferred from the data plotted in Figures 13 and 14 by using equations (43) and (53).



$$\rho_{rms}\ (\%)\ =\ 6.45 + 0.454\ \sin\left[\frac{2\pi}{365}(d-43)\right] + 1.87\ \sin\left[\frac{4\pi}{365}(d+59)\right]$$

d = day number

Figure 16 Seasonal variation of the rms atmosphere density perturbations at Urbana, IL. The data were obtained in 1986-89. The lines are the range of values measured during the night and the dots denote the average value for the night. The solid curve is the least-squares annual plus semi-annual variation.

IX. CONCLUSIONS

Rayleigh and Na lidars are perhaps best suited for studying the vertical and temporal structure of the middle atmosphere. Modern Na lidars can achieve vertical resolutions of a few tens of meters with temporal resolutions of a few tens of seconds. Although the signal levels for Rayleigh lidars are weaker, in the upper stratosphere modern Rayleigh lidars can achieve vertical and temporal resolutions of a few hundred meters and a few minutes respectively. Lidars are attractive remote sensing instruments because they are portable and don't require extensive support facilities for their operation. In fact, airborne Na lidar measurements have been reported [Kwon et al., 1989], and several spaceborne systems have been proposed [Yeh and Browell, 1982; Philbrick et al., 1985]. Spaceborne systems are especially attractive because they can provide global measurements of gravity wave and tidal activity and can be used to study the horizontal structure of these important wave phenomena. Currently there is a significant lack of observational data on the climatology and horizontal characteristics of both gravity waves and tides. In addition very little is known about the geographical distribution and characteristics of gravity wave sources. Airborne and satellite based Rayleigh and Na lidars can help provide that information.

Although most of our observational knowledge of middle atmosphere winds has been obtained with radars, only lidars can provide high resolution measurements of the atmospheric density and temperature. Rayleigh and Na lidars are mature techniques that are being used routinely to observe these parameters in the upper stratosphere, mesosphere and lower thermosphere. In this paper we have reviewed the capabilities Rayleigh and Na lidar systems for measuring temperature, density and wind perturbations and the spatial and temporal spectra of these important parameters. The results clearly demonstrate the potential of these two lidar techniques and illustrate the critical importance of laser power and telescope area in determining the quality of the observations.

Acknowledgements. The authors gratefully acknowledge the many helpful discussions with colleagues C. H. Liu and S. F. Franke at the University of Illinois, C. Y. She at Colorado State University, J. W. Meriwether at the Geophysics Laboratory and C. R. Philbrick at Pennsylvania State University. Some of the data presented here were obtained during the AIDA Act 89 Campaign at Arecibo. We thank Colin Hines for inviting us to participate in this campaign as part of the NSF CEDAR Program. We also acknowledge the excellent support and assistance provided by C. A. Tepley at Arecibo. We thank Nina Parsons for the superb job she did in preparing the manuscript. This work was supported in part by grants from the National Science Foundation and the Air Force Geophysics Laboratory.

REFERENCES

Batista, P. P., B. R. Clemesha, D. M. Simonich and V. W. J. H. Kirchhoff, Tidal oscillations in the atmospheric sodium layer, *J. Geophys. Res.*, *90*, 3881-3888, 1985.

Beatty, T. J., R. E. Bills, K. H. Kwon and C. S. Gardner, CEDAR lidar observations of sporadic Na layers at Urbana, Illinois, *Geophys. Res. Lett.*, *15*, 1137-1140, 1988.

Chanin, M., and A. Hauchecorne, Lidar observation of gravity and tidal waves in the stratosphere and mesosphere, *J. Geophys. Res.*, *86*, 9715-9721, 1981.

Dewan, E. M., and R. E. Good, Saturation and the "universal" spectrum for vertical profiles of horizontal scalar winds in the atmosphere, *J. Geophys. Res.*, *91*, 2742-2748, 1986.

Elterman, L. B., The measurement of stratospheric density distribution with the search light technique, *J. Geophys. Res.*, *56*, 509-520, 1951.

Elterman, L. B., A series of stratospheric temperature profiles obtained with the searchlight technique, *J. Geophys. Res.*, *58*, 519-530, 1953.

Elterman, L. B., Seasonal trends of temperature, density and pressure to 67.6 km obtained with the searchlight probing technique, *J. Geophys. Res.*, *59*, 351-358, 1954.

Fricke, K. H., and U. von Zahn, Mesopause temperature derived from probing the hyperfine structure of the D_2 resonance line of sodium by lidar, *J. Atmos. Terr. Phys.*, *47*, 499-512, 1985.

Fritts, D. C., Gravity wave saturation in the middle atmosphere: A review of theory and observations, *Rev. Geophys.*, *22*, 275-308, 1984.

Gardner, C. S., M. S. Miller and C. H. Liu, Rayleigh lidar observations of gravity wave activity in the upper stratosphere at Urbana, Illinois, *J. Atmos. Sci.*, *46*, 1838-1854, 1989.

Gardner, C. S., and J. D. Shelton, Spatial and temporal filtering technique for processing lidar photocount data, *Opt. Lett.*, *6*, 174-176, 1981.

Gardner, C. S., and J. D. Shelton, Density response of neutral atmospheric layers to gravity wave perturbations, *J. Geophys. Res.*, *90*, 1745-1754, 1985.

Gardner, C. S., and D. G. Voelz, Lidar Studies of the nighttime sodium layer over Urbana, Illinois 2. Gravity waves, *J. Geophys. Res.*, *92*, 4673-4694, 1987.

Gardner, C. S., D. G. Voelz, C. F. Sechrist, Jr., and A. C. Segal, Lidar Studies of the nighttime sodium layer over Urbana, Illinois, 1, Seasonal and nocturnal variations, *J. Geophys. Res.*, *91*, 13659-13673, 1986.

Gibson, A., L. Thomas and S. Bhattachacharyya, Lidar observations of the ground-state hyperfine structure of sodium and of temperatures in the upper atmosphere, *Nature*, *281*, 131-132, 1979.

Hirota, L., Climatology of gravity waves in the middle atmosphere, *J. Atmos. Terr. Phys.*, *46*, 767-773, 1984.

Iwasaka, Y., T. Ono, and A. Nomura, Changes in aerosol content and temperature in the Antarctic spring stratosphere; Lidar measurement at Syowa Station (69°00'S, 39°35'E) in 1983, 1984, and 1985, *Geophys. Res. Lett.*, *13*, 1407-1410, 1986.

Kent, G. S., L. R. Poole, and M. P. McCormick, Characteristics of arctic polar stratospheric clouds as measured by airborne lidar, *J. Atmos. Sci.*, *43*, 2148-2161, 1986.

Kwon, K. H., C. S. Gardner, D. C. Senft, F. L. Roesler, and J. Harlander, Daytime lidar measurements of tidal winds in the mesospheric sodium layer at Urbana, Illinois, *J. Geophys. Res.*, *92*, 8781-8786, 1987.

Kwon, K. H., D. C. Senft, and C. S. Gardner, Airborne sodium lidar observations of horizontal and vertical wavenumber spectra of mesopause density and wind perturbations, *J. Geophys. Res.*, in press, 1989.

Lindzen, R. S., Turbulence and stress owing to gravity wave and tidal breakdown, *J. Geophys. Res.*, *86*, 9707-9714, 1981.

McCormick, M. P., T. J. Swissler, W. H. Fuller, W. H. Hunt, and M. T. Osborn, Airborne and groundbased lidar measurements of the El Chichon stratospheric aerosol from 90°N to 56°S, *Geof. Inst.*, *23*, 198-221, 1984.

Neuber, R., P. von der Gathen, and U. von Zahn, Altitude and temperature of the mesopause at 69°N latitude in winter, *J. Geophys. Res.*, *93*, 11,093-11,101, 1988.

Philbrick, C. R., J. L. Bufton, and C. S. Gardner, A solid state tunable laser for resonance measurements of atmospheric sodium, in *Tunable Solid State Lasers for Remote Sensing, Ser. Opt. Sci.*, vol. 51, R. L. Byer, E. K. Gustafson, and R. Trebino, Eds. New York, NY: Springer-Verlag, 1985, pp. 42-46.

Philbrick, C. R., et al., Density and temperature structure over northern Europe, *J. Atmos. Terr. Phys.*, *47*, 159, 1985.

Reiter, R., and H. Jager, Results of 8-year continuous measurements of aerosol profiles in the stratosphere with discussion of the importance of stratospheric aerosols to an estimate of global climates, *Meterol. Atmos. Phys.*, *35*, 19-48, 1986.

She, C. Y., R. E. Bills, H. Latifi, J. R. Yu, B. J. Alvarez II and C. S. Gardner, Two-frequency lidar technique for mesospheric Na temperature measurements, *Geophys. Res. Lett.*, in press, 1989.

Simonich, D. M., and B. R. Clemesha, Resonant extinction of the lidar returns from the alkalai metal layers in the upper atmosphere, *Appl. Opt.*, *22*, 1387-1389, 1983.

Sivjee, G. G., and R. M. Hamwey, Temperature and chemistry of the polar mesopause OH, *J. Geophys. Res.*, *92*, 4663, 1987.

Smith, S. A., D. C. Fritts, and T. E. VanZandt, Evidence of a saturated spectrum of atmospheric gravity waves, *J. Atmos. Sci.*, *44*, 1404-1410, 1987.

Theon, J. S., et al., The mean observed structure and circulation of the stratosphere and mesosphere, NASA Tech. Rep. TR-375, Mar. 1972.

Vincent, R. A., Gravity wave motions in the mesosphere, *J. Atmos. Terr. Phys.*, *46*, 119-128, 1984.

Vincent, R. A., and D. A. Fritts, A climatology of gravity wave motions in the mesopause region at Adelaide, Australia, *J. Atmos. Sci.*, *44*, 748-760, 1987.

von Zahn, U., K. H. Fricke, R. Gerndt and T. Blix, Mesopheric temperatures and the OH layer height as derived from ground-based lidar and OH spectrometry, *J. Atmos. Terr. Phys.*, *49*, 863-869, 1987.

Welsh, B. M., and C. S. Gardner, Nonlinear resonant absorption effects on the design of resonance fluorescence lidars and laser guide stars, *Appl. Opt.*, *28*, 4141-4153, 1989.

Yeh, S., and E. V. Browell, Shuttle lidar resonance fluorescence investigations, 1, Analysis of Na and K measurements, *Appl. Opt.*, *21*, 2365-2372, 1982.

Chapter 7

Upper Atmospheric Wind and Temperature Measurements using Imaging Fabry-Perot Interferometers

David Rees, Ian McWhirter, Anasuya Aruliah and Simon Batten,

Atmospheric Physics Laboratory,
Department of Physics and Astronomy,
University College London,
Gower Street, London WC1E 6BT, UK

Abstract.

The Fabry-Perot interferometer is a high resolution spectrometer which has a relatively large etendue for a given instrument size and cost, and has thus been very widely used for aeronomic studies, particularly those involving visible and near-IR airglow and auroral emissions. The performance of the classical FPI, however, has been quite critically limited by the normal use of a 'pinhole' detector, namely the photomultiplier. The development of two-dimensional Imaging Photon Detectors provided a very effective means of considerably increasing the sensitivity, thus improving the performance, and of simultaneously reducing the complexity of a Fabry-Perot Interferometer (FPI). These gains have been particularly important for instruments located at remote stations, making upper atmospheric wind and related measurements. Two forms of this Imaging FPI have been developed and deployed by the Atmospheric Physics Laboratory at University College London. The standard Imaging FPI, of which some 10 instruments are now in routine operation at 7 (mainly high-latitude) stations, has a 15 cm clear aperture etalon. One complete Fabry-Perot fringe is continuously imaged and recorded, and the viewing direction is scanned around the sky, through 8 discrete directions, including zenith. The Imaging FPI system has a significant multiplex advantage over non-imaging FPIs (where the optical path difference is scanned, and a pinhole sampler is used), since the entire available signal is recorded all the time. In the Doppler Imaging System (DIS), an all-sky optical imaging system is used ahead of the FPI, to avoid entirely the requirement of a physical scanning system. In the DIS, typically 5 fringes are imaged onto the 2-dimensional detector, further increasing the multiplex gain, while providing a full and continuous 2-dimensional representation of a large upper atmospheric region (400 km radius for the thermospheric OI red line) centred on the observing station. In excess of 2000 nights of useful thermospheric data have been obtained with this class of ground-based instrument since late 1980. The latest improvement with the systems involves the deployment of Imaging Photon Detectors with Gallium Arsenide photocathodes. A recently-deployed system with such a detector has opened up the near-I.R. to FPI observations, initially using the OH Mienel emissions near 840 nm for mesospheric wind and temperature measurements. Some exciting possibilities of imaging two-dimensional wind and temperature fields in the mesosphere, and in the lower and upper thermosphere, will be realised as these GaAs IPDs become more widely available. Some of the highlights of the observations from the Imaging FPIs and DIS will be used to illustrate the capabilities of the instruments.

NS 1.20:W 89/V.2

WORLD IONOSPHERE/THERMOSPHERE
STUDY... VOLUME 2...
NATIONAL SCIENCE FOUNDATION

3 OF 5
24X
SUPTDOCS/GPO
P.0189-0285

1. Introduction

The conventional Fabry-Perot interferometer, as used for ground-based studies of upper atmospheric winds and temperature, has used a pinhole detector to detect the optical signal, while the Optical Path Difference (OPD) of the etalon is changed by pressure-scanning (change of refractive index) the etalon cavity, or by mechanical scanning of the etalon plate separation. These instruments have been widely used since the 1960's (Armstrong 1968), and the methodology is widely reported in the literature, and reviewed by Hernandez (1986), Meriwether (1984, 1989) and Meriwether et al., (1986).

A fundamental limitation of the conventional FPI is the severe restriction on the acceptable solid angle, i.e. that which can be transmitted to a detector without degrading the spectral transmission of the instrument (the Jacquinot criterion). A number of attempts have been made, using photographic methods, image intensifiers, multiple element masks etc, to 'field-widen' the FPI, thus increasing the entendue and thus throughput. However, these were never entirely successful in the context of an 'observatory class instrument', since they usually demanded a considerable effort to build and align the instrument, and a great deal of effort to calibrate, maintain stability, and then analyse the data. Several of the methods which reached the stage of being initially deployed in the field, but not of becoming widely and reliably used, were reviewed by Hernandez (1986).

Two fundamental requirements had to be defined and then solved experimentally in the development of a truly successful instrument capable of overcoming these limitations to the FPI.

Firstly, it required the development of a relatively cheap, compact, simple and reliable two-dimensional Imaging Photon Detector (IPD) of high optical performance and excellent medium and long-term image and photometric stability. A 5-year development programme at UCL, started in 1975, produced the first IPDs which met these requirements in 1979, although very considerable technical and performance improvements have been made since then (McWhirter et al., 1982; Rees et al., 1985, 1986, 1989). One of the major, unsung, advantages of the IPD over other available imaging detectors is the extremely low noise associated with the devices. At the very low detectable signal levels which characterise all aeronomic interferometric measurements, it is not so much the detective quantum efficiency itself which limits performance, but a complex ratio of DQE to read-out and thermal electron noise. This factor will not be discussed in detail here, however, the absence of any read-out noise for the IPD, coupled with a thermal electron noise of 0.01 photon-equivalent event per 60 sec per pixel, and no susceptibility to cosmic ray events ensures that it is only the 'square root "n"' of the detected signal itself which contributes to limiting the instrument performance under given very low signal level conditions. This factor is of extreme importance for determining limiting performance / signal levels for airglow studies, in particular.

Secondly, it required the development of a very stable and rugged Fabry-Perot etalon, including the necessary mounting and other structures, so that a non-scanned imaging instrument could be built. The key factor here, which otherwise may not be obvious, is that the regular scanning of the OPD does provide a regular auto-calibration of the FPI. In a non-scanning, imaging system, other methods of calibration of the OPD and other parameters have to be developed and utilised. The Fabry-Perot etalon technology was developed in conjunction with I.C. Optical Systems of London, in a programme starting in 1970 / 71. This programme was aimed at the development of rocket and satellite-borne FPIs, and was exploited with several successful rocket flights of UCL FPIs in the mid 1970s and early 1980s; in a collaboration with the University of Michigan in which UCL developed and provided the flight etalons of the DE-2 Fabry-Perot interferometer; in two high altitude balloon flights of the UCL triple-etalon FPI (1980 and 1983) which demonstrated the technique of stratospheric wind measurements from interferometric observations of molecular absorption-lines; and in the development of prototypes of etalons for the High Resolution Doppler Interferometer of the NASA Upper Atmosphere Research Satellite (Rees 1979; Kaskiewicz 1978; Rees et al., 1982; Killeen et al., 1982; Rees et al., 1983, 1984).

There was also a third, implicit, limitation before the mid 1970s, associated with on-line data handling. The two-dimensional detectors provide a challenge for a non computer-based signal processing system,

BEST COPY AVAILABLE

in terms of electronic complexity, development time and cost. The advent of machines based on the 8080 / 8085, and comparable microprocessor chips revolutionised the design of the signal processing system required for the Imaging Fabry-Perot interferometers. By 1980/81, software to run these machines had been implemented on 8085 and 6502-based personal computers, and on DEC 11-23 machines. These machines, for the first time, allowed the software and straightforward serial / parallel interfaces to carry out all the functions of instrument control, timing, data input, processing, display and storage.

The rapid advent of personal computers, during 1981/2/3, provided another major boost to the cost-effectiveness of these instruments. PCs provide an immediate solution encompassing a convenient programming environment, very cost-effective computational power and data storage, flexible interfacing, the capability of auto-start and re-boot using back-up internal clocks and inbuilt reliable hard disk facilities for the operating system and software and for data storage. The global availability of systems, servicing and spares also completely changed the nature of the limitations to the sophistication of the overall instrument and its operating system: since the computer replaced a very complex and by its nature relatively unreliable and difficult to repair controller system, it rapidly became possible to build a very simple and reliable controller and power supply unit, set through its many functions by a very cheap and reliable personal computer. With regular data back-up, even a complete disk crash, nearly the worst scenario, can be solved for all but one of our instruments by a visit to the local and friendly personal computer supermarket. For the outstanding instrument, at Halley Bay, Antarctica, the only solution to problems is still an abundant supply of spare parts and instruction manuals. The very considerable simplification of the electronics, and the easy general availability of the PC made the 'cloning' of such instruments an attractive possibility, when coupled with the demonstrated high performance and combination of reliability and ruggedness in operation.

The Imaging Fabry-Perot interferometers (FPI) have been widely used to study the motion of the middle and upper thermosphere by measuring Doppler shifts of the OI 630.0 nm emission (Rees et al., 1984, 1985, 1987; Smith et al., 1985; Meriwether et al., 1986; Stewart et al., 1988). Similarly, neutral winds in the lower thermosphere may be derived, under the proper observing conditions (in particular, lack of auroral contamination), from Doppler shifts of the OI 557.7 nm emission. Such measurements, on a routine basis, have been extremely useful for describing the neutral dynamics within the thermosphere. Both middle / upper thermosphere and lower thermosphere observations provide a useful input for the thermospheric general circulation models (TGCM). They also supply an independent measurement of the winds in the lower thermosphere for comparison with the data from incoherent and partial reflection radars.

There have been previous attempts at interferometric Doppler wind studies of the OH Mienel emissions near 840 nm. The early versions of the IPD had photocathodes characterised by S-20 response, with the occasional device reaching the S-25 specification (extended red response). Due to the relatively high cost of the IPD (the 'cheap' being relative to similar detectors for Astronomy and Space Science applications), it was, until recently, necessary to run them in a conservative way, with a significant reduction in Detective Quantum efficiency. Lately, devices with a full S-25 photocathode specification have become available from ITT (USA). ITT have also started production (currently very limited numbers) of IPDs with Gallium Arsenide (GaAs) windows. These latter devices, in particular, have completely revolutionised interferometric studies of near IR airglow emissions such as OH. Some examples of recent very successful OH studies, and thus of winds near 85 km altitude, will be discussed later.

This first exploration of the region around 850 nm will be followed by further forays, using ground-based instruments, to study even weaker emissions, well below previously achievable thresholds. One of the limitations from the ground is, however, the relatively high near IR background night-glow continuum (or quasi-continuum), mainly of atmospheric origin. From space-borne instruments observing the atmosphere at the limb, the background intensity level is usually much lower. The intensity of specific emission features is considerably enhanced by the Van Rhijn effect. Thus there are considerable possibilities for future development and exploitation of these techniques and instruments. While space-borne instruments have the greatest potential, via their capability for global coverage, the number of future space-borne instruments will be quite limited due to the high cost and limited number of suitable space platforms. Ground-based instruments will thus continue to have a significant role in the future investigation of upper atmospheric dynamics, in particular.

2: Instrument Description.

I: The Imaging Fabry-Perot Interferometer.

The group of Imaging Fabry-Perot Interferometers developed by the University College London Atmospheric Physics Laboratory is primarily intended to measure upper atmospheric winds via the Doppler shift of the same airglow and auroral emissions as those used by conventional Ground-based Fabry-Perot interferometer instruments (i.e. Meriwether, 1989). The major distinction has been the exploitation of a two-dimensional Imaging Photon Detector (McWhirter et al., 1982), to take advantage of a multiplex gain, reading the entire Fabry-Perot Fringe pattern continuously, rather than sequentially stepping through the spectrum by changing the Optical Path Difference (OPD), and thus being limited to deploying only a small, on-axis, sampling aperture.

The complete instrument is illustrated schematically in Figure 1. The FPI comprises a scanning mirror system, the F-P etalon, a focussing telescope, an interference filter, the detector and a spectral calibration source. These components are mounted from a rigid and rugged four-bar optical bench, to ensure that the alignment and focus of the entire system is maintained. The entire instrument, sealed by light-tight covers, is normally mounted within an enclosure which is dark and light-tight and thermally controlled to within about 3 - 5° C. As for most of these FPI instruments, a plexiglass dome provides access to the night sky. The system is completed by an electronic controller unit (ECU), run from a fairly standard IBM PC (XT or upward in specification), the detector signal processing unit (SPU), and a thermal control system for the detector. The latter is normally a power supply for the Peltier cooler built into the detector, and a cooler / pump for removing heat from the hot junction of the Peltier elements via a liquid (water) system.

The etalons of the Imaging FPIs (for the OI 557 or 630 nm emission) have a nominal 14 mm gap and a 15 cm diameter clear, working, aperture. The three spacers of the etalon are made of zerodur, so that the thermal expansion coefficient of the etalon is of the order of 0.2 - 0.4 $10^{-6}K^{-1}$. The etalon itself, mounted within a sealed and evacuated container, is temperature-stablised to +/- O.1 K, so that the Optical Path Difference is not affected by changes of ambient air pressure and refractive index. Further details of the etalon design and other instrument details can be found in Rees et al. (1982) and Rees et al. (1984). The appropriate spectral region required for a specific investigation is normally selected with a 1 nm interference filter, to reduce unwanted contamination from night-sky continuum, or from other nearby airglow or auroral line or band emissions. Usually, about 1.0 to 1.5 fringes, corresponding to a field of view of about 1.0°, are imaged by a telescope (lens or Cassegrain) of suitable focal length onto the photocathode of the Imaging Photon Detector. The IPD photocathode has a diameter of either 18 or 25 mm. The detector uses a resistive anode as the position-sensing technique, described elsewhere by Rees et al., (1980, 1981) and McWhirter et al, (1982). For each detected photon, the detector's output signals are converted by analogue computation into a 256 by 256 digital (x,y) image coordinate representation. These data are stored in computer memory, where the full 2-D image is normally built up, with a 16 bit counter allocation for each pixel in the image array. The nominal wavelength resolution of one image pixel is wavelength and instrument-dependent, but is of the order of 0.2 pm. This translates, typically, to an equivalent Doppler shift of around 100 ms^{-1}. To maintain instrument stability the temperature of the etalon container is actively controlled to better than 0.1 Celsius, at a temperature a few degrees above the highest anticipated observatory temperature.

In recently-built FPIs, detectors of high red-sensitivity using Gallium Arsenide (GaAs) photocathodes have been used to provide a greatly improved response to the OI 630 nm emission, and to those emissions such as OII 730 nm and OH 840 nm which are even further into the near IR. It is essential to cool the photocathodes of such detectors to -20C or -30C, to reduce the thermionic emission to a negligible level. A Peltier cooler is used, with the hot junction being cooled by a circulating water system. Depending on the requirements, this circulating water system is itself cooled or thermally

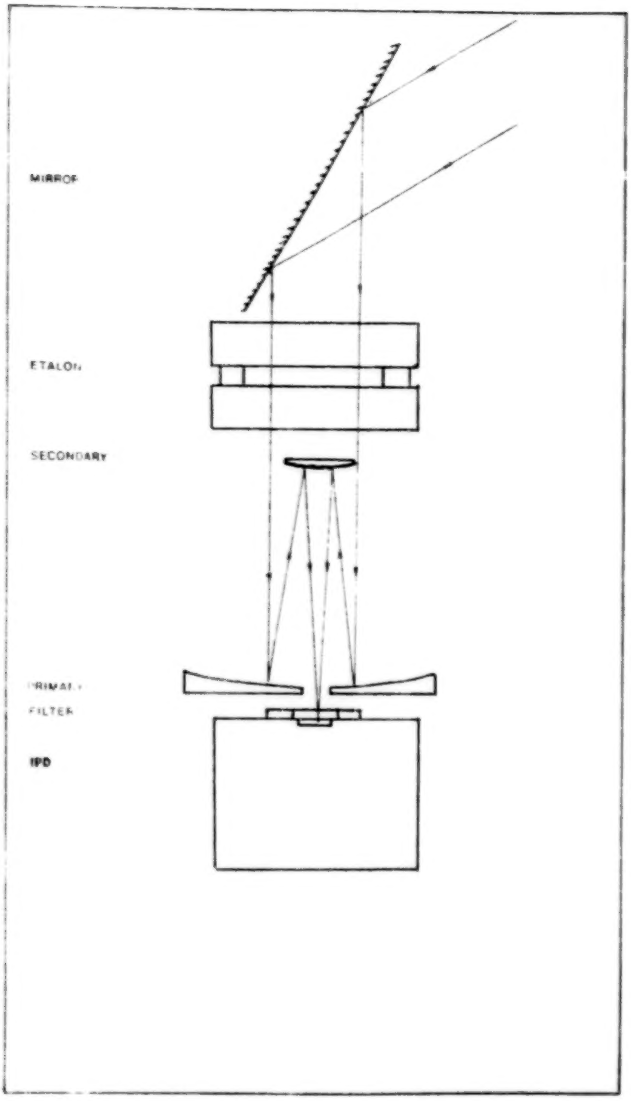


Figure 1: A schematic diagram of the Imaging Fabry-Perot Interferometer, showing how the component parts are arranged within the instrument, and the path of light rays through the instrument.

controlled, to maintain the stability of the detector photocathode temperature. This Peltier cooler is specially designed for the format of the particular IPD. In practice, the water-backed systems have proven rather more reliable than air-cooled versions. Ensuring the combination of a good air circulation and zero light leakage is a particular problem with air-cooled Peltier systems.

II: The Doppler Imaging System.

The UCL Doppler Imaging System is presently operating from the optical station at the Swedish Institute for Space Studies (IRF). It is a Fabry-Perot interferometer of high spectroscopic resolution with two major modifications. Firstly, a high resolution, two-dimensional, Imaging Photon Detector (IPD), (McWhirter et al., 1982) in the focal plane records the images of five individual Fabry-Perot fringes. Secondly, a wide-angle (all-sky) optical system is placed ahead of the Fabry-Perot etalon, to match a large field of view in the sky to the field of view of the Fabry-Perot fringes in the focal plane. In practice, about 100° arc (full angle) in the sky corresponds to about 2.5° arc (full angle) in the image plane, and the 100° corresponds to the centre of the 5th Fabry-Perot fringe. Individual angular and radial segments of each Fabry-Perot fringes are thus coherently mapped onto the sky. Given a knowledge of the height of a specific emission layer (and assuming that the height is approximately uniform), the subsequent data analysis provides a map of line-of-sight (l-o-s) wind velocity and relative auroral intensity over the entire field of view. A subsequent data analysis stage is required to reconstitute the (assumed horizontal) vector wind field.

Figure 2 shows the schematic layout of the DIS, with the arrangement of optical components, and the light path through the instrument. The general principles of the optical system of the DIS and the necessary calibration procedures to ensure reliable and accurate data have been discussed previously in Rees and Greenaway (1983), Rees et al. (1984) and Batten et al. (1988). Many of the practical aspects of the DIS are identical to those of the UCL FPIs. However, the simplification involved in removing the scanning mirror system does introduce some complications of the calibration system in addition to those required of the FPIs, while removing one source of moving parts and inherent potential unreliability.

Briefly, the following calibrations of the DIS have to be performed:

I) Flat field sensitivity calibration, to correct for non-uniformity of optical throughput and detector sensitivity over the image plane.

II) Thermionic emission calibration, to correct for thermal electron emission, and its variation over the detector surface.

III) Full-field spectrometric calibration, using the neon 630.4 nm line, to map the base-line Optical Path Difference (OPD) variations over the image plane.

IV) Positional geometrical calibration, to map sky locations to position on the detector image plane.

V) Continuous 'spot' calibration, to monitor thermal, electronic and mechanical instrument drift.

The DIS shares the useful property of the FPI that the entire spectral pattern is recorded simultaneously. Rapid temporal fluctuations of intensity over the field of view are generally of less importance than with a spectrometer or interferometer which has to be sequentially scanned in OPD. Since the sky is imaged directly onto the photocathode of the IPD, we have always been aware that sharp spatial features and intensity gradients in the far field are imaged directly onto the Fabry-Perot fringe pattern. With the FPI, due to the small instantaneous field of view, the upper atmosphere winds invariably advect any significant airglow or auroral intensity features over the image plane within the integration period. However, for the OI 630 nm emission, such gradients are normally very small, in any event. For the DIS, with its large field of view, there is a greater potential problem caused by intensity gradients. In practice, for studies using the OI 630 nm emission, there is never a practical problem. Advective smearing due to winds, and the low spatial intensity gradients of the OI 630 nm emission always cause the intensity modulation of the fringes to be below a critical level at which wind

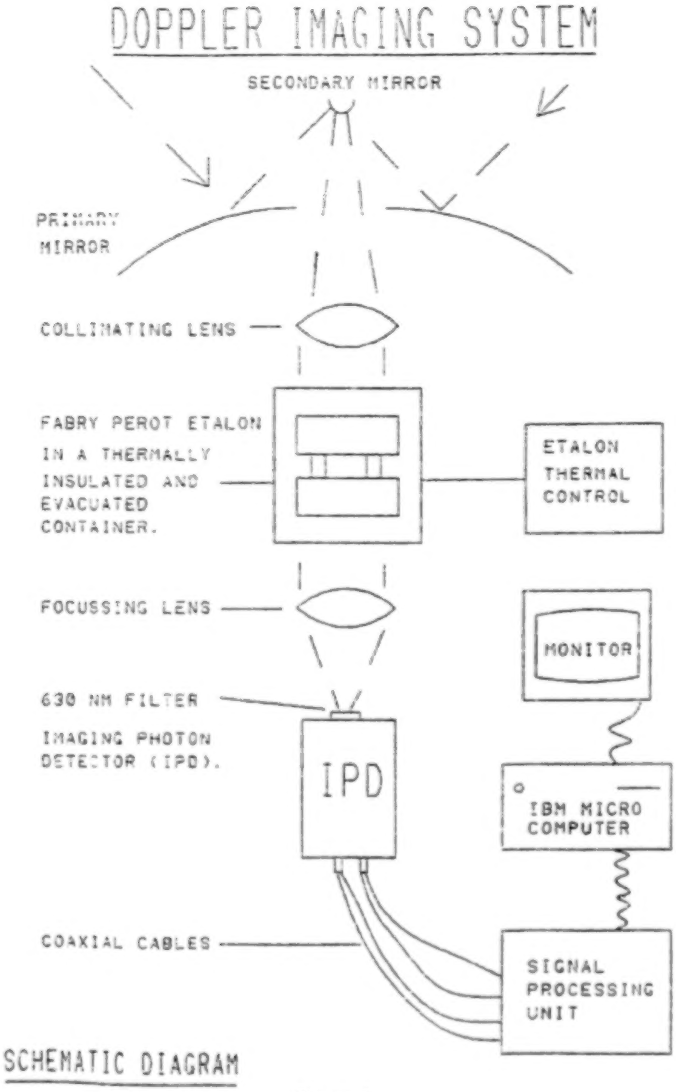


Figure 2

Figure 2: A schematic diagram of the DIS showing how the component parts are arranged within the instrument, and the path of light rays through the instrument.

errors would occur. We should, however, note that with lower thermosphere or mesospheric applications, the intensity gradients are larger, and the advective smearing is lower. Methods which have been developed at UCL to analyse DIS observations of rocket-borne shaped-charge releases (which have enormous and rapidly-changing gradients over the fringe pattern) may then have to be applied to correctly analyse and interpret the DIS fringe data.

3. Instrument Calibrations and Data Analysis:

I. The Imaging Fabry-Perot Interferometer.

The raw Fabry-Perot image data, of images of 256 by 256 pixel are accumulated into the computer memory (Figure 3). The two dimensional images, in normal operation mode, are compressed for storage, into a one dimensional array (Figure 4), comprising a radial profile - in effect the spectrum. The centre of the Fabry-Perot fringe has to be determined to the nearest pixel, before this 'reduction to radius squared' operation is performed, automatically, at the end of each integration period. Under normal observing conditions, only these one dimensional profiles and spectra are stored onto computer disk. Doppler shifts are inferred from the spectral analysis of the fringes. The radial profiles are corrected with the appropriate thermionic emission and flat-field response prior to spectral analysis.

Further corrections and calibrations may have to be applied to the radial profiles, prior to spectral analysis. Firstly, any non-linearity in the detector geometric response may be deconvolved. This correction ensures that the apparent distance, in wavelength space, between the calibration line and the zero Doppler shift of the observed auroral or airglow emission, is a constant of the instrument, rather than a possible variable. Although the instruments are, in principle, extremely stable in operation, the normal intervention of Mr. Murphy dictates that regular spectral calibration should be used to monitor changes of OPD. A neon (or other suitable gas such as Krypton) R-F excited lamp is the normal regular calibration source, with the illumination of a screen viewed by the FPI each rotation around, normally 8 discrete viewing positions. The occasional failure of a thermal control system, opening of external doors, miniscule leaks of the vacuum seal of the etalon container can thus be coped with by the regular calibrations, without any loss of data. Otherwise, it is often several months between a problem occurring and its being noted in data returned to the UK for analysis. Typically, it may be a full year before a minor problem can be fully rectified. Thus the regular calibrations provide the means of determining the health status of the instrument. Many minor problems can be corrected using the calibration data, avoiding the loss of data. Frequently, early warning signs of potential problems can be noted, improving the efficiency of planning for the infrequent but expensive visits for preventative maintenance, and sometime avoiding potentially catastrophic problems which might develop later.

The regular calibrations also allow comparison between measurements recorded with different instruments, and provide continuity between data sets obtained after minor or major changes in FPI / IPD configuration, and between instruments at different locations.

When fully corrected, each spectral profile is then analysed to determine the fringe radius, line intensity, line width and continuum intensity. There are several possible methods of analysing these spectral profiles, including that implemented by Killeen and Hays (1984) for the 12-segment multi-ring anode Image Plane Detector. Methods which have been used for peak-fitting have been reviewed by Lloyd (1984) and by Aruliah (1990). The analysis includes allowance for the calibrated instrument profile, determined periodically by a laser calibration source. This latter feature is essential for determination of temperature from the observations.

One of the most difficult factors to ascertain is the true effective baseline for wind absolute determination. Two factors need to be determined: the absolute zero velocity baseline, and its relative changes from time to time. In practical terms, it is possible to routinely and regularly determine the offset, in wavelength space, from the neon calibration lamp fringe at 630.4 nm to OI 630 nm line. For studies of the OI 557.7 nm emission, the position of a convenient fringe from a Kr calibration lamp is used to monitor relative drifts of the 'zero Doppler shift'. Other calibration lamp emission lines are available close to most of the other airglow and auroral emissions which are interesting targets for

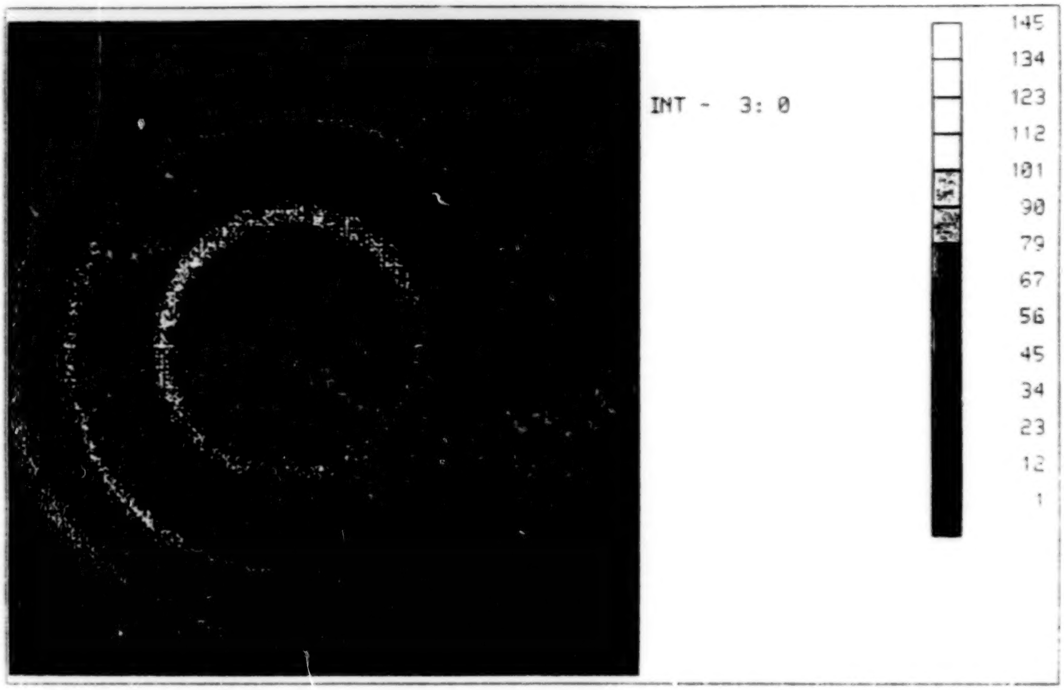


Figure 3: The raw Fabry-Perot image data, of 256 by 256 pixel, as accumulated into the computer memory.

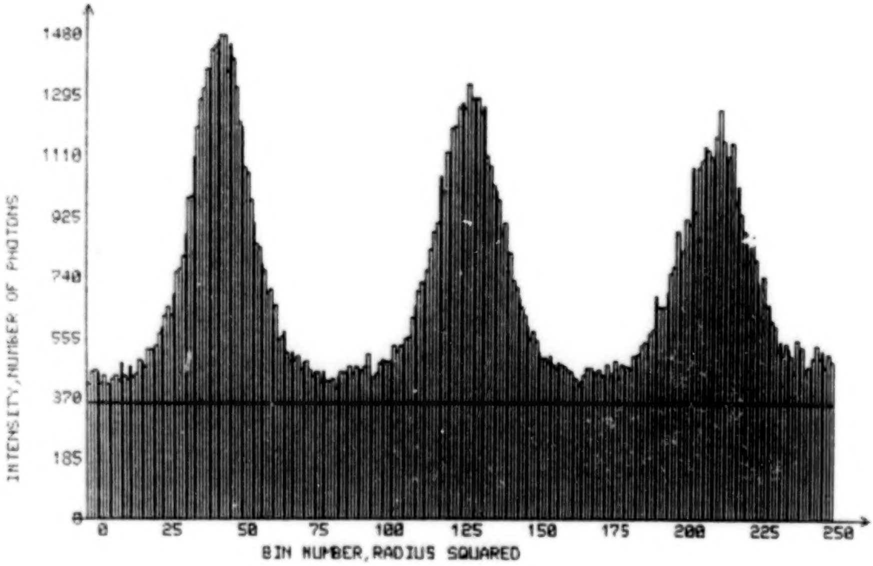


Figure 4: The two dimensional images, in normal operation mode, are compressed for storage, into a one dimensional array comprising a radial profile - in effect the spectrum.

study. The key factor is that the calibration emission can be transmitted through the interference filter used to select the atmospheric line being studied, so that there is absolutely no possible change of the spectral characteristics of the instrument from operating to calibrating mode.

Given this relatively easy method of deducing and monitoring relative drift, the more difficult problem of determining the absolute zero velocity baseline has to be faced.

The normal method is to use all available and valid zenith observations taken over an extended period (several months) of observations, corrected for relative drifts by using the R-F excited spectral lamp calibrations. The usual assumption is made that, when averaged over a long time period, the mean vertical wind would be small, almost certainly less than 10 ms^{-1}. Much larger vertical winds occur for relatively short periods, and particularly at the high latitude stations (Rees et al., 1984a,b). The validity of this assumption has to be cross-checked each night by ensuring that the zenith Doppler shift lay midway between the mean Doppler shifts obtained from opposing cardinal directions. In short, establishing that the instrument, data and data analysis methods are all working reasonably, the long-term base-line can be established to within 10 ms^{-1}. Given this base-line, relative changes are then monitored using the R-F calibration source. In principle, a convenient laboratory source of the OI 630 nm line would remove some residual ambiguity, and such sources have been demonstrated to operate for a moderate lifetime. However, the R-F lamps have a very long lifetime and are both cheap and reliable, and are also very easy to replace within the instrument. Those sources of OI 630 nm emission which have been demonstrated still have some drawbacks in terms of limited lifetime and a need, because of relatively low intensity, to arrange a very efficient optical illumination system into the interferometer. Regrettably, this would involve considerable re-design and engineering of the present existing instruments.

There are several methods of presenting the data obtained from individual measurements of the Doppler shift obtained in the series of directions. Firstly, neglecting the vertical wind (which can be, and usually is observed, but only directly overhead the observing station), the line-of-sight wind component can be converted into an equivalent horizontal wind component. The sequence of wind components obtained in successive directions can then be used either independently, or alternatively, they may be combined. For example, the mean horizontal wind flow can be estimated from the least squares fit of a cosine to observations of the six cardinal, line-of-sight, Doppler shifts. It is then possible to analyse for a linear gradient in the wind flow. The FPI observing circle is normally relatively large, spanning some 4^{o} - 8^{o} of latitude, depending on the zenith distance observed, and the altitude of the emission layer. Values for the individual wind components, related to their well-separated locations around the 'observing circle', plus the mean wind and linear gradient can be tabulated either 'as obtained', or alternatively at 15 minute intervals by linearly interpolating between individual data points. It should be noted that the mean wind is relatively insensitive to any error in the point of zero Doppler shift as it mostly depends upon the difference between Doppler shifts from opposing directions. At high latitude stations in the auroral oval or polar cap, the gradients and temporal changes of winds are often very large, particularly under disturbed geomagnetic conditions. The mean winds have then to be referred back to the 'absolute zero velocity baseline'. Frequently, for detailed interpretation, for example comparison with model simulations, only the independent wind components are directly useful for other than statistical analysis.

Under some conditions, a calibrated photometer has been located alongside the FPI to provide direct calibration of the FPI line and continuum intensity measurements. Under these conditions, the FPI line intensities have been converted to an equivalent zenith intensity (i.e. Lloyd et al., 1989). More usually, an all-sky camera is run along-side (normally part of the standard instrumentation at sites such as IRF, Kiruna). These all-sky data provide the necessary monitor of poor or variable observing conditions. A meridian-scanning photometer has also been used from time to time to provide a means of cross-calibrating the overall sensitivity of the Imaging FPI.

All Fabry-Perot wind measurements are susceptible to corruption from atmospheric scattering due to aerosols and clouds. Normally, the entire data set is inspected to locate periods when scattering may have been significant. Scattering, by any mechanism, is usually characterized by a reduced mean Doppler wind and may be accompanied by large and sudden apparent wind gradients and changes, as for example if a patch of cloud suddenly moves into the field of view. Under very cloudy conditions, the Doppler shifts and thus 'winds' measured in all directions normally fall to very low values, and there is an accompanying reduction in fringe contrast which almost uniquely identifies the period of next to useless data. Conversely, the detection of a large mean wind with small gradients implies that scattering is probably insignificant. Clouds are an obvious source of scattering, but low altitude aerosols [Abreu et al., 1983], local atmospheric pollution and the observing dome are other potential scattering sources. Clearly, in studying many phenomena, long-term, unattended observation is essential, and it is quite impractical and prohibitively expensive to always have an operator in attendance merely to record sky and viewing conditions. Paradoxically, light pollution from nearby cities creates the fringe contrast reduction which usually paints marginal or variable conditions in an unambiguous way.

The effects of tropospheric scattering are most severe when there are large intensity gradients across the FPI observing region in the sky. This is often the case during periods of strong auroral activity. Results taken during such periods do occasionally, display the characteristic signature of scattering, as for example simulated by Abreu et al. (1983) and thus may have been contaminated. Generally, such periods have been excluded from the final data set. This is particularly important with long-term studies, and with statistical analyses using data from extended periods.

The instruments are normally attended for approximately 1 month, each instrument, each year, when it is possible to fully characterise the instrument, the data, and the response to specific meteorological conditions. These periods of attendance have been used to build up an extended data base of information relating data quality to observing conditions (and to particular airglow and auroral activity). This data base from observations with an operator in attendance amounts to approximately 10 % of the total data, or at least 300 nights out of the total of more than 2000 nights of useful data, and is the basis for selection of nights, observing directions and observing periods when the data is believed to be of adequate quality for analysis and interpretation.

II. The Doppler Imaging System.

The Doppler Imaging system shares much common hardware and software with the Imaging Fabry-Perot interferometer. It is not surprising, therefore, that the calibration procedures (tabulated earlier) and the data software also have a common body of methods and analytical routines.

Figure 5 shows a full two-dimensional image of the sky obtained by the DIS during a period of moderately active auroral conditions at 06:13 UT on 27th November 1987. This image is the raw data obtained by the DIS. North is directed upward, east to the right. In the southern part of the image, part of a distinct second set of fringes can be seen. These are from a neon calibration lamp (630.4 nm), illuminating a calibration 'patch' on the plexiglass dome covering the DIS. There is a dark area in the centre of the image. This is the shadow of the secondary mirror holder. The three dark spokes in the image are the out-of-focus silhouette of the secondary mirror spider mount. The central shadow masks the first FP fringe, while the outermost fringe is incomplete and partly vignetted, and cannot normally be used in the wind analysis.

The first stage in the analysis of the DIS data is the real-time reduction of the full 2-D images to 24 spectral profiles. The photon counts in 24 equal area sectors, each of 15° arc, are reduced to a linear wavelength presentation, based on the centre of the Fabry-Perot fringe pattern. This process is exactly equivalent to that shown in Figure 4 in respect of the Imaging FPI. An example of this reduction is illustrated in Figure 5 of Batten et al. (1988). The centre of the fringe pattern must be accurately determined before any reduction takes place - this is a process which is performed, usually by iteration, using a spectrum obtained from the R-F excited neon calibration lamp. The 24 spectral profiles - 'reduced to radius squared' - due to the quadratic relationship between radial location and wavelength in the image plane, are then analysed a-postiori to provide:

(i) fringe peak locations with error estimates,
(ii) a map of the intensity of the OI 630.0 nm emission (96 elements), and
(iii) a map of the background sky intensity (96 elements).

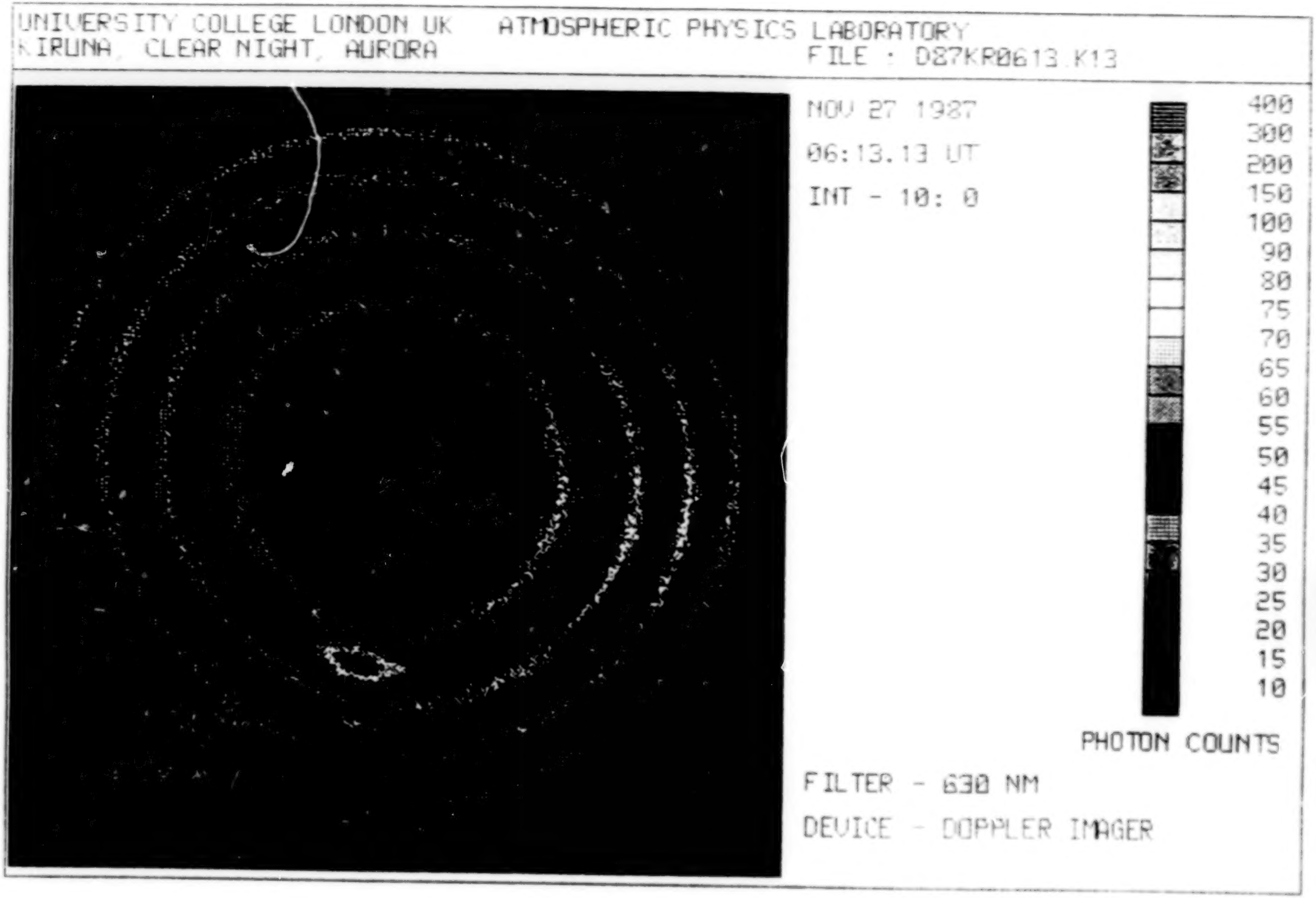


Figure 5: A full two-dimensional image of the sky obtained by the DIS under conditions of moderately active auroral activity at 06.13 UT on 27th November 1987. North is at the top, east to the right.

After the locations of each of the spectral peaks are determined, the Doppler shifts can then be calculated, and the l-o-s wind velocities over the field of view can be deduced, based on the calibration data, and on a determination of the 'zero velocity' fringe positions. The 'zero velocity' fringe determination has to be based on the analysis of a considerable body of real data, since it is not practical to routinely operate a source of the OI red line at remote field locations. The principle is fundamentally similar to that used for more traditional ground-based FPIs. The analysis of the DIS data to this stage has been fully discussed in Batten et al. (1988).

The absolute line-of-sight wind velocity component (v) is calculated from:

$$v \quad = c \, . \, \lambda / 2t \, . \, (cal - dat) / fsr$$

where:

c = velocity of light in vacuum
λ = wavelength of the emission
t = the etalon gap
cal = position of the calibration peak
dat = position of the data spectral peak after instrument corrections
fsr = free spectral range (adjacent peak positions)

Each typical DIS image is reduced to twenty four 15^0 sectors, each containing four complete fringes which can be fully analysed. This generates 96 individual measurements of the radial l-o-s wind velocity for each integration period. The integration period is usually chosen to be ten minutes, giving six images, and 576 individual l-o-s wind measurements per hour.

4. Vector wind determination for the DIS.

The 96 individual l-o-s wind components are reduced to equivalent horizontal (and radial) wind components by the simple assumption of a zero vertical wind component. The vertical wind directly over Kiruna is monitored by means of the co-located FPI. The RMS vertical wind is normally less than 10 ms^{-1}, however, on occasions, much larger vertical winds do occur (Rees et al., 1984a,b). The assumption of zero vertical wind over the entire field of view of the DIS may, obviously, introduce a significant error under some conditions. The 96 measurements of the horizontal wind component, radial to Kiruna, are then analysed by a maximum entropy method, to calculate the mean wind field.

To produce a unified vector wind field, a continuity condition must be assumed, effectively coupling the data from adjacent azimuthal, radial, and temporal elements of the derived wind patterns. The question of how best to derive the maximum information on the horizontal wind field has to be addressed. There are localized structures within the field, and all the azimuthal, radial, and temporal information available from the individual, and sequential, Doppler images has to be used. A strong wind observed at one location, at one azimuth, should have corresponding features at adjacent azimuth angles, in adjacent fringes, and in consecutive images. It is therefore assumed that the vector wind field is a single valued, continuous, function and that it can be described by a two-dimensional polynomial. It is further assumed that the vector components can each be represented by two orthogonal components and that these components can be described by perpendicular polynomials.

At this stage it is possible to set up the least squares criterion:

$$\chi^2 = \sum_{i=1}^{np} [\, V(r,\Theta) - M(r,\Theta) \cos[\Theta - \phi(r,\Theta)] \,]^2 / \sigma^2 \qquad [1]$$

A Fourier low-pass filter is used to smooth the azimuthal components and to interpolate between missing data points. Weighted Chebyshev interpolation is then used to calculate the vector component of the field in all of the twelve measured, radial, cross-field directions.

The wind vector is described by the equation:

$$W = V\cos(\theta-\psi) \qquad [2]$$

where:

W = line-of-sight wind velocity
V = wind velocity
θ = observing direction
ψ = bearing of wind vector

$$\Rightarrow \quad W = V\cos\theta.\cos\psi + V\sin\theta.\sin\psi \qquad [3]$$

The first harmonic of the Fourier transform is:

$$W = A\cos\theta + B\sin\theta \qquad [4]$$

Therefore, the harmonics A and B can be equated to the average wind velocity V and Bearing ψ by:

$$A = V\cos\psi \quad \text{and} \quad B = V\sin\psi \qquad [5]$$

The direction and amplitude of the true vector (of which A and B are the orthonormal components) can be calculated at each point.

The true wind direction is given by $\psi = \text{Arctan}(B/A)$, and the wind velocity amplitude by $(A^2+B^2)^{1/2}$. Each component of the vector wind field has thus been estimated from all the available l-o-s data, weighted for distance from the point in question. The set of all the individual wind vectors can be plotted to show the structure of the wind field at an altitude assumed to be 240 km. With the given geometry of the all-sky optics of the DIS, this corresponds to a circle of approximate radius of 240 km, centred directly above the Kiruna. The deduced horizontal wind pattern contains the maximum amount of valid spatial information available within each Doppler image, consistent with the quality of the available data. The data quality varies within each image and from image to image, mainly as a function of the localised emission rate for O^{1D}. The χ^2 criteria is observed to decrease steadily as the number of Fourier and polynomial components is increased, until it settles down to a fairly constant value χ_0^2. Once a steady value is reached, this indicates that the closest approximation justified by the data quality has been achieved. If the degree is too high, the value of χ^2 decreases again. This indicates that the data is being over-fitted, and that noise is being included. The order of smoothing also takes account of a maximum entropy constraint.

The maximum entropy method (MEM) is a powerful theoretical approach to the problem of selecting one of a choice of several, otherwise equally valid, solutions to a data set, making the least number of assumptions about unmeasured parameters (Burch et al., 1983; Gull and Daniell 1978; Gull and Skilling 1984; Skilling and Bryan 1984; Sibisi 1983). The maximum entropy method always selects the smoothest solution, i.e. that solution which includes the minimum structure consistent with the data. The entropy, S, is calculated for the magnitudes of the resulting wind fields for different orders of polynomial and Fourier smoothing.

$$\text{Entropy, } S = -\sum_{i=1}^{n} p_i \log p_i \quad \text{where} \quad p_i = f_i \,/\, \Sigma f_i \qquad [6]$$

where f_1 to f_N are a set of numbers, in this case those relating to the wind fields. The chosen orders of smoothing were those which gave the maximum entropy consistent with the constraint of χ^2 approximately equal to χ_0^2.

Figure 6 displays a series of full images of the Doppler horizontal wind components directly derived from the line-of-sight measurements, data obtained between 20.33 UT and 21.38 UT, 25th January 1988. It is only the directly-measured line-of-sight component which is used in this analysis stage, with the calculation of the component of the assumed horizontal wind from the measured l-o-s component. It can be seen that the 'winds' are entirely convergent / divergent - an artifact which is totally unrealistic. The OI 630 nm intensity patterns are displayed as the background scalars / intensity values.

There should be continuity between successive Doppler images. Within the disturbed auroral oval, accelerations of the thermospheric wind can be very large at times. However, within a given gas parcel, only a finite wind velocity change can occur within the ten minute interval between successive images.

Using the analysis method described for deriving the vector wind components from the measured l-o-s components, the corresponding wind vector patterns are displayed in Figure 7. The vector wind patterns, derived from the measured l-o-s components by fitting for the unmeasured radial wind components, as described earlier, appear as a much more realistic field flow pattern. These images were obtained at intervals of ten minutes on this night of moderately disturbed auroral and geomagnetic activity. The sector at 180^{0} azimuth contains no measured wind or intensity information, due to the calibration patch (Figure 5). The wind vectors have, however, been interpolated across this calibration sector by the analysis / fitting process. Between 20:55 and 21:18 UT, there is a general surge in the southwestward wind velocity, particularly noticeable in the southern part of the field of view. This corresponds to the auroral brightening in the southeast part of the field of view of the DIS. In the same period, the winds in the northern part of the field of view slightly decrease, indicating a significant wind acceleration as the thermospheric wind flow crosses the arc. At this stage, it is not possible to determine whether ion drag or the Joule heating associated with the arc play the larger role in the observed wind changes.

There are several inherent problems in the study of small scale wind velocity structures.

(i) Small cloud patches:

Small patches of cloud can cause scattering of the 630.0 nm radiation in a part of the field of view, giving anomalous results, while not warranting the identification or rejection of the whole hourly period as cloudy. Clouds usually produce a localized region of anomalously low wind velocities. This is, usually and fortunately, accompanied by a localised decrease in intensity. The presence of cloud above Kiruna is routinely recorded by an all-sky camera (16 mm colour film). These data are analysed to provide a tabulated hourly average of auroral intensity and weather conditions, published in the Kiruna geophysical data summary on a three monthly basis. Caution is obviously necessary when interpreting data obtained during periods of partial cloud cover, and a review of the all-sky camera film is usually used to verify conditions during periods of specific interest. No data from partly cloudy periods has been included in this present analysis.

(ii) Vertical winds

Vertical winds can further complicate the analysis. The DIS would interpret a uniform upward motion as a divergent wind field. A uniform downward motion would, conversely, appear as a convergent wind field. As yet, no systematic account is made of the possible existence of vertical winds in the analysis of the DIS data. Vertical winds generally have relatively small values compared with the horizontal wind velocities. At times, however, vertical winds of more than 100 ms^{-1} have been observed (Rees et al., 1984a,b). The vertical wind directly over Kiruna is measured routinely by the co-located FPI, however, these local observations are certainly unlikely to be valid for the entire region observed by the DIS.

(iii) Advective Smearing

A parcel of gas moving at 100 ms^{-1} travels 60 km during the typical ten minute integration time used by the DIS. For such an average wind velocity, structures smaller than 60 km scale size will be partly

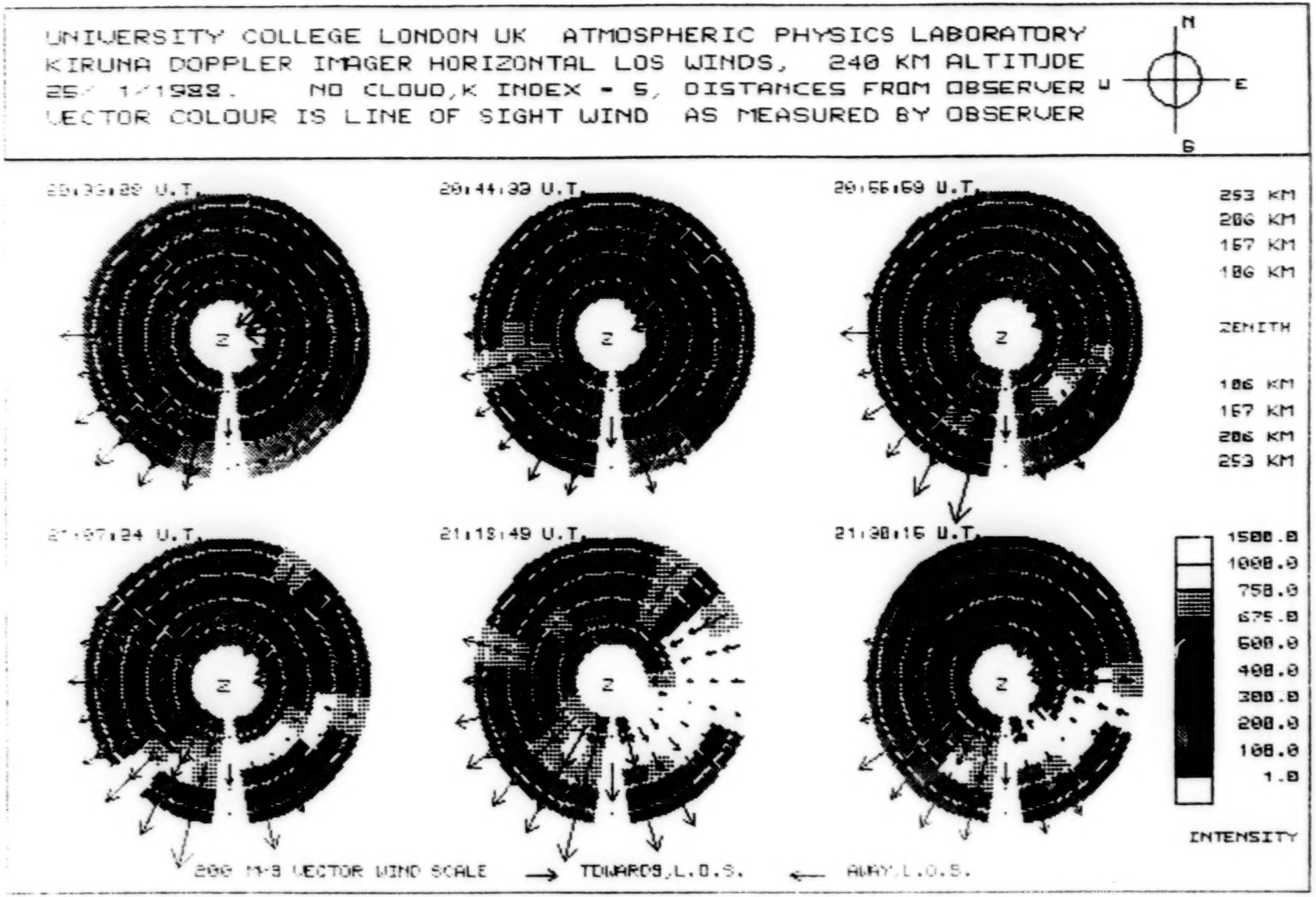


Figure 6: A series of full images are shown of the line-of-sight Doppler shifts, converted into equivalent horizontal wind components obtained between 20.33 UT and 21.38 UT, 25th January 1988.

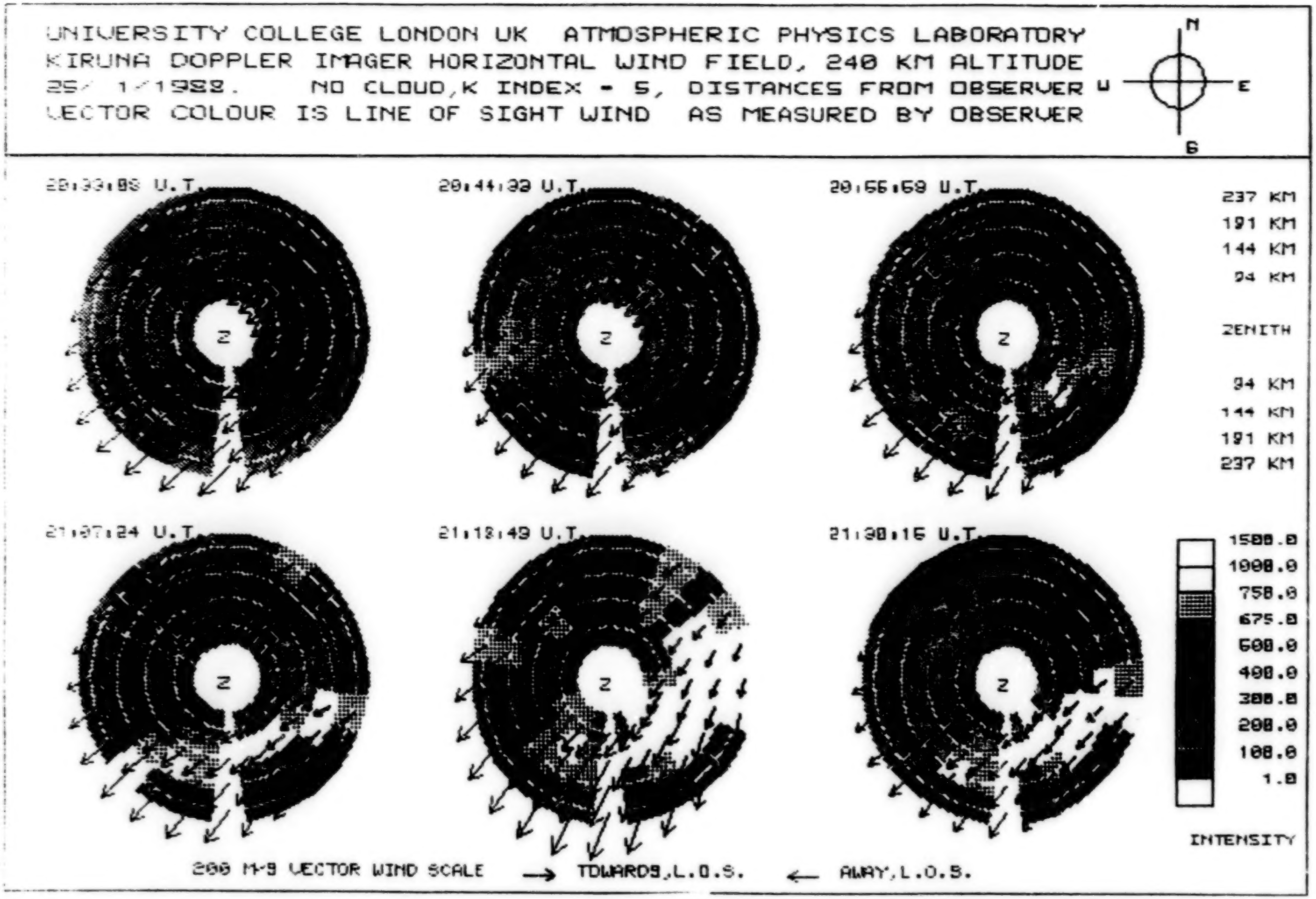


Figure 7: The horizontal vector wind patterns, corresponding to the MEM fitting to the line of sight wind components shown in Figure 6, are displayed.

smeared. The degree of smearing increases with wind velocity: for a 300 ms^{-1} average wind velocity, any structures smaller than 180 km will be partly smeared during a ten minute integration period.

(iv) The maximum entropy technique

The maximum entropy technique, employed to fit the individual l-o-s wind velocity vectors, itself introduces further smoothing. The structures discernible in the data thus tend to have scale sizes significantly larger than the minimum which may actually occur in the thermosphere.

Practical Solutions

Solutions to these problems mainly involve experience with the data handling and common sense. In principle, advective smearing can be reduced by decreasing the integration time. However, during periods of low intensity several successive images would have to be processed to obtain a sufficiently high signal to noise ratio. This solution is thus partly limited by signal levels and detector sensitivity, and partly by the logistics of data storage. Using a ten minute integration time, the DIS currently stores more than a megabyte of data per night. Shorter integration times would require larger magnetic or optical disk storage facilities, both of which are now realisable, but have not yet been implemented. Improvements in the signal to noise ratio can also now be obtained using IPDs with Gallium Arsenide photocathodes. It is hoped to be able to operate the DIS at IRF, Kiruna, with such a device during the 1989 / 1990 winter.

The problems of partial cloud cover and vertical winds require a careful assessment of the available additional information from the all-sky camera and the co-located FPI. Inevitably, the result is that some perfectly valid data is rejected along with data which is certainly contaminated either by local cloud scattering or by localised regions of high vertical winds.

The data shown in Figure 7 (Jan 25, 1988) illustrate the problem of advective smearing. The wind vectors displayed are of the order of 250 to 300 ms^{-1}. There is little discernible small scale structure in the images, and the wind flow appears uniform over the whole period displayed. At 21.18 UT there is an enhancement of the wind flow in the southwest of the image and a general anticlockwise rotation in the eastern half of the field of view. Both of these features could perhaps be related to the enhancement of 630.0 nm emission which started at 21:07 UT, and continued through 21:38 UT.

For lower wind velocities, i.e. less than 100 ms^{-1}, the smearing effect is not so significant and smaller scale structures can generally be seen. However, for low wind velocities, the errors on the individual measurements constitute a larger proportional error than for higher wind velocities.

5. Analysis of Temperature data.

The general principle of temperature measurement with the imaging FPI and DIS systems is identical to that of the conventional FPI (Hernandez, 1986, Meriwether, 1990). The instrument spectral / line profile has to be derived with adequate accuracy, and deconvolved from the observed line profile. This process is included in the routine analysis process for the Imaging FPI and for the DIS. Because of the way in which a number of fringes are compressed onto the imaging detector of the DIS, it is inherently less capable of determining accurate temperatures than is the Imaging FPI. Since, at present all of the Imaging FPIs use high resolution etalons without a facility for scanning the OPD, the process of calibrating the line profile is somewhat more complex than for the conventional FPI, since each part of the free spectral range has to be calibrated for its instrument function. The calibration fringe and the OI 630 nm fringe , for example, rarely overlap in the image plane. This means that an ad-hoc arrangement for pressure scanning has to be introduced to obtain an initial sequence of instrument profile calibrations of the entire detector, before the instrument can be settled into a mode where, by use of a single mode stable laser, any minor changes can be monitored regularly.

We should, however, note that with the Imaging FPI and DIS, there is an important practical aspect, relating to the remote location and automatic operation of all of the FPI instruments of this class. The R-F excited neon lamps are very stable sources of wavelength calibration, and have excellent reliability.

However, they provide a relatively broad spectral line (corresponding to, typically, 700 - 900K). They also do not have a particularly stable line width, since this is dependent on the operating conditions. Single-mode stabilised He-Ne lasers certainly can be used to derive the instrument function with adequate accuracy but are, however, still relatively expensive and unreliable for long-term operations of a remote automated system. Only one of the present Imaging FPI systems (Bear Lake, Hardware Ranch, Utah State University) currently has such a stable laser incorporated into the system for routine calibrations of the instrument profile. This is the only Imaging FPI facility which presently makes routine temperature measurements in addition to the wind flow observations. For some special studies, particular efforts have been made to derive an effective instrument profile from calibrations with an unstabilised laser. Some results comparing thermospheric temperature measurements from the FPI at IRF, Kiruna, and ion temperature measurements from the EISCAT radar were, for example, published by Rees et al (1984c).

6. The Instruments and their Typical Results.

The Imaging Fabry-Perot Interferometers and DIS: Stations, Observing Periods and Data.

The Imaging FPI instruments which have been built at UCL, their locations and normal operational modes, are summarised in this section. These instruments, with the exception of the FPI at Halstead UK, are all operated by the local cooperating research institute. The Halley Bay instrument belongs to the British Antarctic Survey, while that at Saskatoon belongs to the Institute for Space and Atmospheric Studies, University of Saskatchewan, Saskatoon. The following section provides a little more information on each station and instrument. More information, and the detailed scientific results obtained from each of these instruments and stations, can be found in the appropriate references.

Kiruna, Sweden: (67.8°N, 21.2°E)

Since the 1980 / 81 winter, at least one of the imaging FPIs has been operated during the winter period at the Institutet RymdFysik, Kiruna, and since the 1981 / 1982 winter, one or more instruments have been maintained in continuous automatic operation.

At present, there are three imaging FPIs in routine operation at IRF, as well as a Doppler Imaging System. Two of the FPIs are single-etalon instruments, currently making routine observations of the OI red and green line emissions. Another FPI is a triple-etalon system. This instrument is still in a process of development for daytime / year-round thermospheric wind measurements of OI 630 nm. The instrument is based on the hardware used for the UCL high altitude balloon flights (Rees 1985) where its target has been to measure stratospheric wind in daytime by observing the Doppler shifts of absorption lines due to molecular oxygen and water vapour. The triple-etalon system has been used for a series of developmental tests of the triple-etalon imaging FPI concept. It is, however, an expensive and relatively complex instrument. During the 1987 / 1988 period, it was run for 12 months continuously. It is expected that in late 1989 or early 1990, the instrument will be further upgraded with two new high-performance capacitance-stabilised etalons, and with a new 32-channel, multi-element anode detector recently obtained from ITT.

Some 40 papers have been published which have involved the results obtained from the UCL FPIs at Kiruna, including the use of their data in conjunction with large-scale observations from Dynamics Explorer, the global network of Fabry-Perot interferometers and incoherent scatter radars (such as within the SCOSTEP WITS and GITCAD / LTCS programmes). There have also been a number of studies using combined data sets from the nearby EISCAT radar facility.

Figures 8 and 9 show some particularly interesting storm-time results from the peak of the last solar cycle. In Figure 8, the behaviour of the thermospheric wind in the afternoon hours during a disturbed period is highlighted (Dec. 12, 1981). Between 14 and 16 UT, the westward (sunward) winds, excited initially by ion drag within the expanded auroral oval, reached speeds in excess of 900 ms^{-1}. This situation has been discussed in some detail by Fuller-Rowell and Rees (1984), and corresponds to resonant flow within the dusk auroral oval and polar cap regions. In Figure 9, the comparable situation

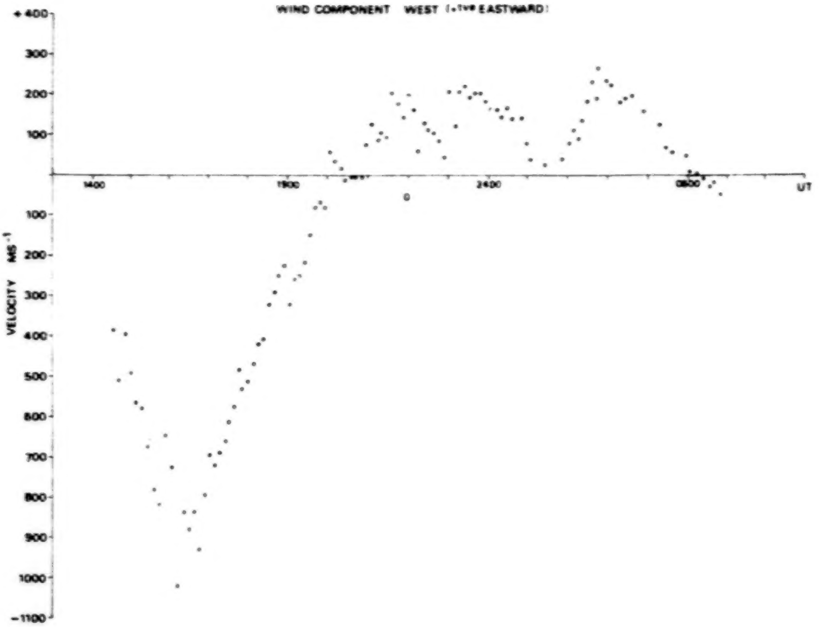


Figure 8: The response of the zonal component of the thermospheric wind at Kiruna during a disturbed period 14 - 19 UT on Dec. 12, 1981.

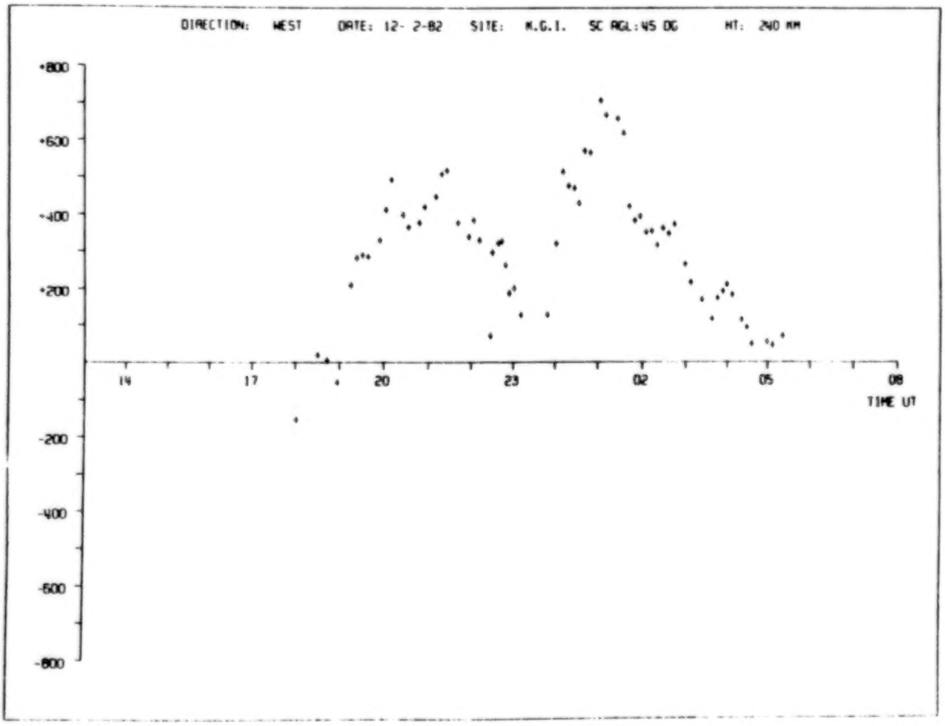


Figure 9: The response of the zonal component of the thermospheric wind at Kiruna during a disturbed period 19 - 03 UT on Feb. 12/13, 1982.

in the dawn auroral oval is shown. Normally, in this region, it is quite difficult to generate very strong sunward winds via ion drag. Feb. 12 1982 was, however, an unusually disturbed period, with the total IMF reaching 40nT. In addition, there was a strong negative IMF BY component (Rees et al, 1986). This condition strongly enhances ion flow in the dawn auroral oval, and leads to a situation where the eastward (sunward) wind components peaking at approximately 700 ms^{-1} could be generated and maintained for several hours. There were also extreme (500 - 600 ms^{-1}) southward winds during this unusual period. This particular disturbed period continued for approximately 4 days.

More than 1000 nights of high quality data are available, obtained during clear sky conditions, from the Kiruna instruments. The statistical analysis of these data have been presented and discussed by Lloyd (1985), Rees et al. (1985, 1987) and by Aruliah et al. (1989).

Halstead, England: (0.1°E, 51.3°N)

The instrument operated here is the 'home base' (literally in the 'back yard' of D.R.). Given the famous average UK weather, this station is generally used for development and testing work, shaking out the variety of bugs in new instruments before they are dispersed, unattended, to the far parts of the world. The instruments which are set up from time to time are also used for training purposes, to reduce the initial problems associated with encounters between students and relatively inexperienced or new staff and the instruments. Various problems which occur in instruments in remote places can sometimes be duplicated, and occasionally solved by attempting solutions on a local instrument, otherwise nearly identical to the instruments out in the field. In 1981-1983, the effects of several large geomagnetic disturbances were recorded. The data obtained under these conditions were qualitatively similar to those discussed later under Utah State University Instrument. Similarly, at the present time, 1989, two large disturbances (March 1989, Sept. 1989) have been observed.

Halley Bay, Antarctica: (75.5°S, 26.8°W, L = 4.2)

This instrument was built in 1981/82 for a collaboration between Dr. R.W Smith (then at Ulster Polytechnic), the British Antarctic Survey (Dr. Alan Rodger, BAS, Madingley Road, Cambridge, UK) and UCL. The instrument has been in regular operation at the BAS station at Halley Bay in the Austral Winter months most years since then. It is by far the most remote of the Imaging FPIs in this group, and requires a greater amount of instrumentation back-up than any other system, to overcome the various problems. Very low humidity, and the once-per-annum visit which is possible for significant re-supply and major maintenance are among the greatest problems. Nevertheless, the instrument has run reliably for most of the period after the first couple of observing winters. The instrument provided the first southern polar observations of thermospheric winds from a ground-based FPI. Until very recently, the only other south polar thermospheric wind available was that from the DE-2 satellite-borne FPI.

Figures 10 and 11 show some typical results from the Halley Bay instrument in the 1987 austral winter, which are discussed in detail by Stewart et al (1985). In Figure 10, the winds from a relatively disturbed day (12 / 13 July 1983) are shown, compared with a global numerical model simulation. In the early morning hours, equatorward winds reach 200 m/sec, very similar to values in the model, although the observed winds between 08 and 10 UT differ significantly from the model prediction of a continued equatorward component. In Figure 11, the wind response observed on a quiet day (July 15 / 16 1983) are shown. In this case, the maximum equatorward winds are significantly lower in the period from midnight to 06 UT. Other results have been presented by Stewart et al. (1988).

Svalbard, (Initially at Nye Alesund, later at Longyearbyen): (78.2°N, 15.6°E)

UCL have operated, in collaboration with the University of Tromso (Dr. Kjell Hendricksen), one of the FPIs at a station in Svalbard since 1983/4. Again, the remote nature of the site and the low humidity have created operational problems which needed to be overcome before reliable operation could be maintained. Since 1985, the station has operated most of the time during most winters, however, it is the least attended and supervised of any of the UCL instruments, so that even the slightest problem can cause a partial or complete loss of data.

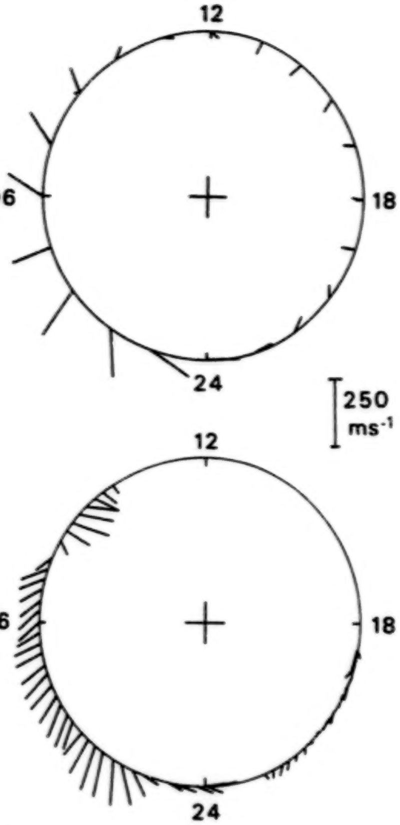


Figure 10: Thermospheric wind data from the Halley Bay instrument in the 1987 austral winter. July 12 /13 1983 was a disturbed day, with Kp varying between 3+ and 5, and with prolonged local magnetic activity. The predicted winds from a model simulation (UCL / Sheffield Model) are in the upper segment, while the observed winds are in the lower segment.

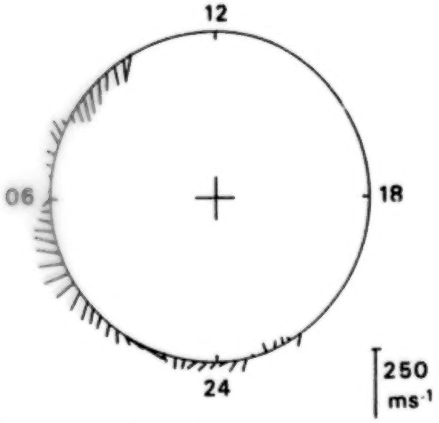


Figure 11: Thermospheric wind data from the Halley Bay instrument in the 1987 austral winter. July 15 / 16 1983 was a geomagnetically quiet day.

In conjunction with the data from the conventional FPI operated there by the University of Alaska (Dr. R.W. Smith), thermospheric wind data is available from Svalbard for most of the last solar cycle. The station shares an advantage with a few other stations: during the midwinter period, the instrument can operate for some two months on a continuous basis, in a period when the sun is always more than 6 degrees below the horizon. There is a distinct advantage in obtaining 24 hours (and sometimes several days) of continuous, unbroken, data about thermospheric circulation. This is of great advantage in comparisons with model predictions, and provides continuity of phenomena and particular events which is often lost when the lower-latitude instruments switch on or off in the middle of interesting events.

Some typical results from the Svalbard instrument are shown in Figures 12 and 13, representing relatively quiet, and very disturbed geomagnetic conditions, respectively. On the quiet day, 27 December, 1987, the wind magnitudes are less than 200 ms^{-1}, with a general pattern corresponding to a flow directed anti-sunward. On February 12, 1988, as the result of the strong geomagnetic activity, the winds are generally 300 - 500 ms^{-1}. In the midnight period, the winds are strongly equatorward (300 - 500 ms^{-1}), while the zonal component changes from eastward before UT midnight to westward after UT midnight. Typical zonal velocities on this day are 300 - 400 ms^{-1}. Again the flow is that within the geomagnetic polar cap, the strong flows responding mainly to ion drag forcing within the disturbed auroral oval, in response to disturbances in the solar wind and a large (80-100 Kv) cross-polar cap electric potential.

Kilpisjarvi: (68°N, 21°E)

Since 1982 / 1983 winter, a 15 cm instrument has been operated at Kilpisjarvi, Finland, in collaboration with the Finnish Meteorological Institute (Dr. Risto Pellinen, Helskinki, and Sodankylia Observatory). The instrument is located at the home of the late Mr. Harinkowski, whose enthusiasm and continual support were the mainstay of the operation at Kilpisjarvi for many years.

The instrument is relatively remote from the point of view of maintenance, however, the instrument operations are regularly supervised, so that minor problems (full data disks, snow on the dome etc) are noted and solved rapidly. The instrument has been used to monitor the OI red line 630 nm most of the time. At that location, it is a little more in the 'centroid' of the EISCAT observing region, and also of the STARE radar system. Additionally, given the local variations of weather during the Scandinavian winter period, a second site, although only 160 km north of Kiruna, is of great assistance in providing continuity, some data cross-check, and extending the periods of joint observations, as with EISCAT, when detailed studies of thermosphere-ionosphere interactions can be studied.

Occasionally, the thermospheric wind data from Kilpisjarvi are available when all three of the northern European stations are clear. Under those conditions, the combination of the data from all three stations provide a very interesting insight into the behaviour of thermospheric flows under various geomagnetic activity conditions, and through the dusk, midnight and dawn periods. Three examples of data obtained under such conditions are shown in Figures 14, 15 and 16. The continuity of thermospheric wind flow between the three stations is quite clear. In Figure 14, the flow reversal occurs between Svalbard and Kilpisjarvi during the afternoon hours 15 to 20 UT. In the midnight period (20 to 02 UT), there is a continuous flow over the polar cap, into the auroral oval stations, with some divergence of the flow to the south of the Kiruna station. In the early morning hours, there are strong southwestward winds observed at Svalbard, but these flows do not reach Kilpisjarvi and Kiruna, where the flow becomes changeable around the dawn period. These data were obtained during relatively quiet to moderately disturbed conditions.

In Figure 15, the winds observed at Svalbard are very strongly eastward during the period from 18 to 23 UT. The wind flow early in this period at Kilpisjarvi and Kiruna is relatively weak and changeable, but becomes southeastward after 18 UT. There is no clear flow reversal between Svalbard and Kilpisjarvi on this night. During the midnight hours there is a lack of flow continuity between Svalbard and the other stations, and again the flow patterns are variable during the dawn period, when the auroral oval passes poleward of Kiruna and Kilpisjarvi. The overall flow distribution during this period was quite

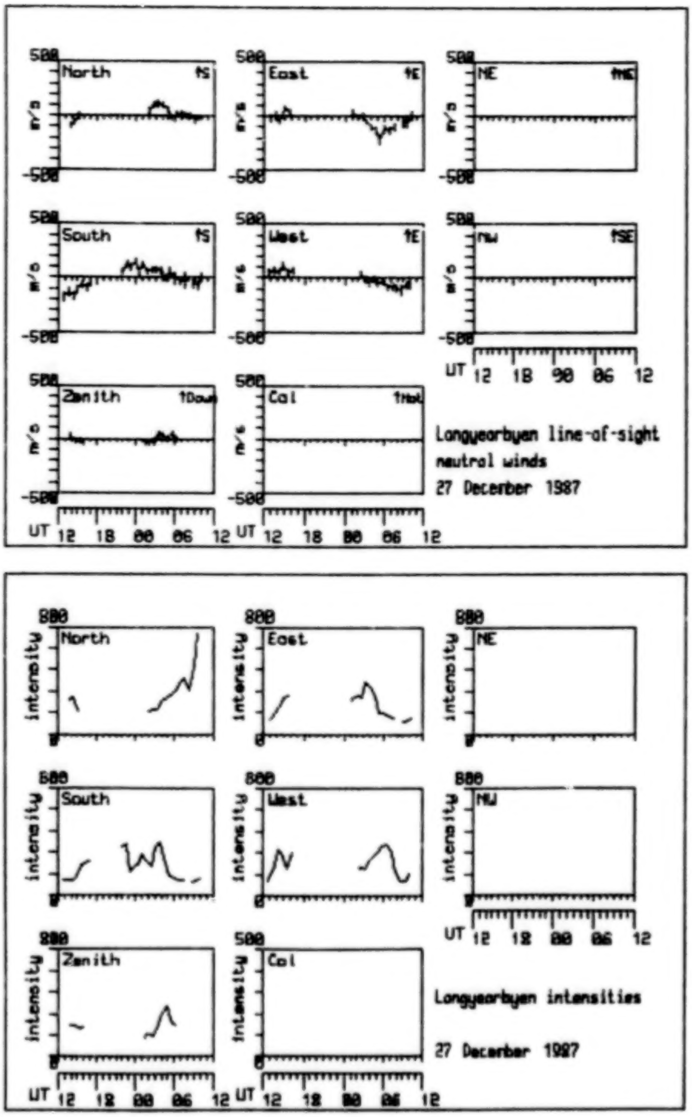


Figure 12. Thermospheric wind data from the Svalbard instrument on December 27, 1987, under relatively quiet geomagnetic activity conditions.

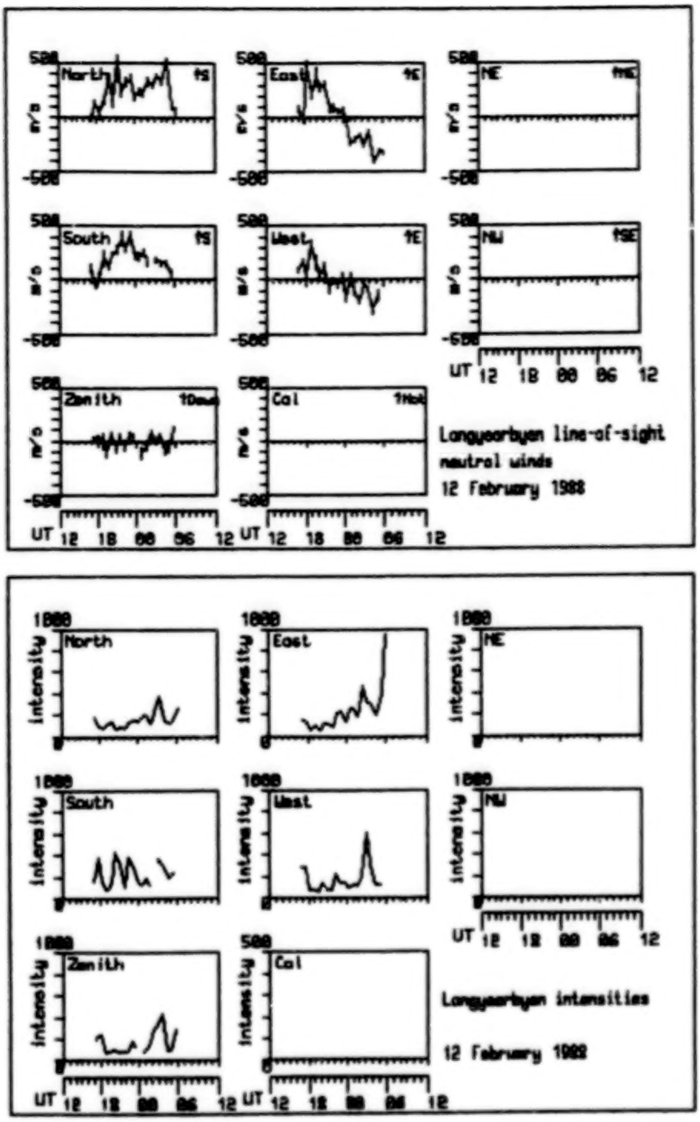


Figure 13. Thermospheric wind data from the Svalbard instrument on February 12, 1988, under very disturbed geomagnetic activity conditions.

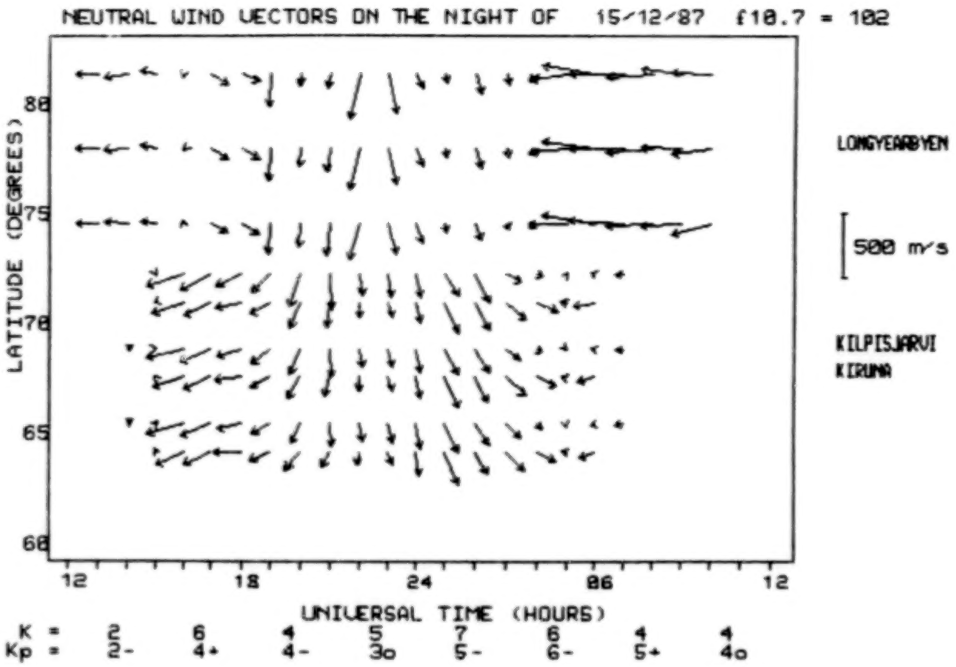


Figure 14. Thermospheric winds from three stations - Svalbard, Kilpisjarvi and Kiruna obtained on December 15 / 16, 1987, a moderately disturbed day.

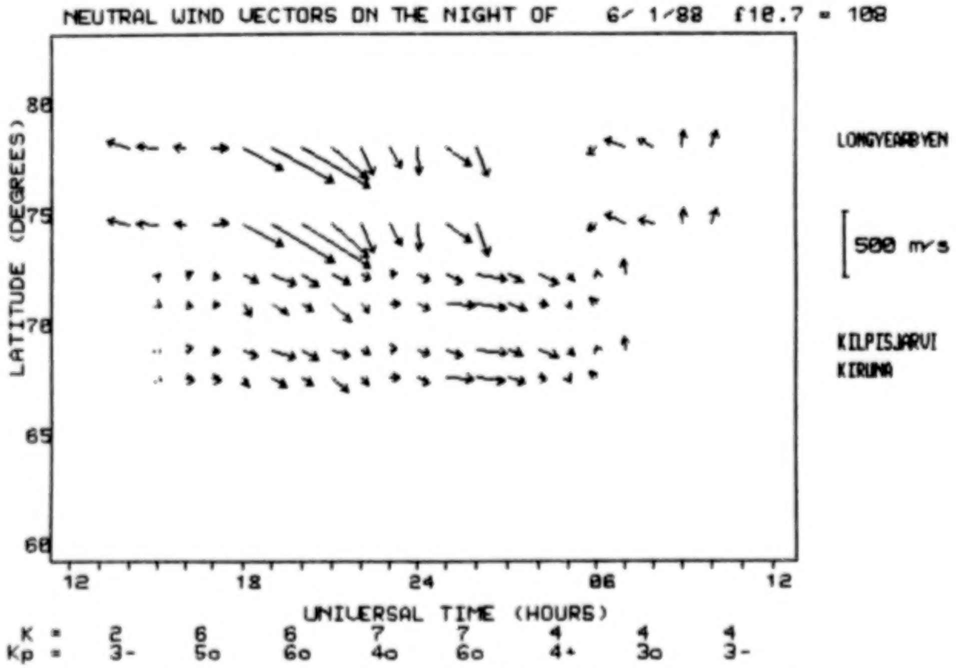


Figure 15. Thermospheric winds from three stations - Svalbard, Kilpisjarvi and Kiruna obtained on January 6 / 7 1988, a moderately disturbed day.

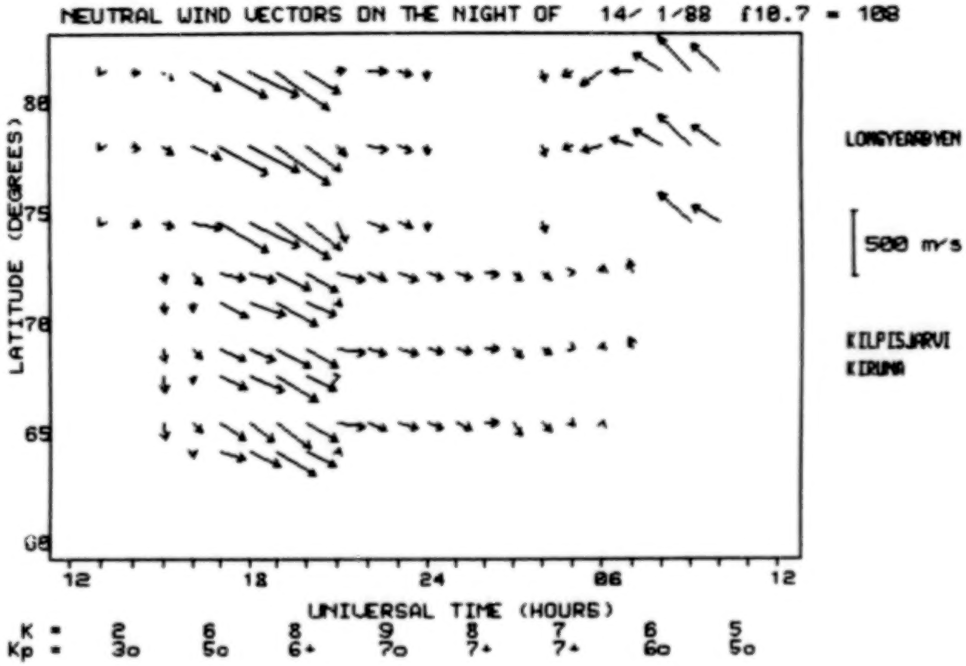


Figure 16. Thermospheric winds from three stations - Svalbard, Kilpisjarvi and Kiruna obtained on January 14 / 15, 1988. This was a very disturbed period, at a time when the auroral oval had expanded equatorward of Kiruna, even before the start of the observing period in the early afternoon.

unusual by virtue of the relative lack of coherence between the polar cap (Svalbard) and the two, relatively close (1000 km), auroral oval stations.

The data shown in Figure 16 was obtained during a very disturbed period, when the auroral oval had expanded equatorward of Kiruna before about 16 UT, in the early afternoon. The winds at all three stations are indicating very strong anti-sunward flow within a greatly-expanded polar cap. The response after 23 UT, at the height of the disturbance, and the recovery from this very disturbed period after about 05 UT were not observed so clearly due to a period of obscured sky at Svalbard. These data were also obtained during an unusual disturbance, and indicate the thermospheric response to a very high level of energy and momentum forcing within the greatly expanded auroral oval and polar cap.

Saskatoon: (52°N, 107°W, L=4.3)

The Imaging FPI at Saskatoon was built for Prof. Don McEwen (Institute for Space and Atmospheric Studies), and has been routinely operated there by Dr. N.D. Lloyd. The Fabry-Perot interferometer at Saskatoon has measured 557.7 nm Doppler shifts since August 1986 and Doppler wind flow data on some thirty six clear nights, from October 1986 to June 1987, have been published (Lloyd et al., 1989). These data corresponded to the period near solar minimum, and presented an opportunity to study the quiet thermosphere.

Utah State University: (42°N, 110°W)

During mid 1988, an Imaging FPI belonging to UCL was installed at Center for Atmospheric and Space Science (CASS) at Utah State University. Dr. Robert Sica and Dr. Vince Wickwar are the local scientists who have been responsible for all of the accommodation and installation arrangements. Initially, the instrument was operated from the roof of CASS, close to down-town Logan, studying OI 630 nm, while awaiting a GaAs IPD from ITT. That device was installed in August / September 1989. Since then, the instrument has been moved to a dark-sky site at Hardware Ranch / Bear Lake, and the instrument is now routinely observing mesospheric winds using the OH Mienel emissions at 843/845 nm. Initial data from this system are presented in Figures 17 and 18.

In Figure 17, the OI 630 nm winds are displayed for a very disturbed night, August 29, 1989. In the local midnight period, the equatorward wind reaches 500 ms^{-1}, with a 200 - 300 ms^{-1} westward component as well. Considering the geomagnetic location of the USU station, the wind response is due to ion drag and heating in a greatly expanded auroral oval, with a very large cross-polar cap potential. The auroral oval did not expand quite as far equatorward as the FPI station, but bright aurora were observed in the northern part of the sky for much of the period of FPI observations. These data were obtained with the FPI immediately (2 hrs) after installation of the GaAs IPD.

In Figure 18, some data obtained on the OH 843/845 nm emissions with the new GaAs IPD are presented (Rees et al., 1989). The wind scale for these data is 50 ms^{-1}, rather than the 500 ms^{-1} used to present the thermospheric wind results. Although only some 5 nights of data on OH winds have yet been analysed, the mesospheric wind flows (around 85 km altitude) present a pattern of considerable day to day regularity, indicating tidal amplitudes of the order of 30 - 50 ms^{-1}. These can be readily seen in Figure 18. There are some very interesting day to day variations, and times when on particular nights, short term fluctuations of winds and intensity occur. It was not possible to run the stabilised single-mode He-Ne laser during this period of observations, however, when that is commissioned in the very near future, the regular calibration of the instrument function will permit mesospheric temperatures to be derived, in addition to sustaining the study of the upper mesospheric winds, of which these are the first results.

The Doppler Imaging System: (IRF, 67° N, 22°E)

The DIS is located at Kiruna, and has been described previously. In its initial configuration, the data handling and data analysis operations were extremely burdensome. Recently, with a dedicated PC, and large disk, plus some improved methods of calibration and data analysis, it has been possible to run the

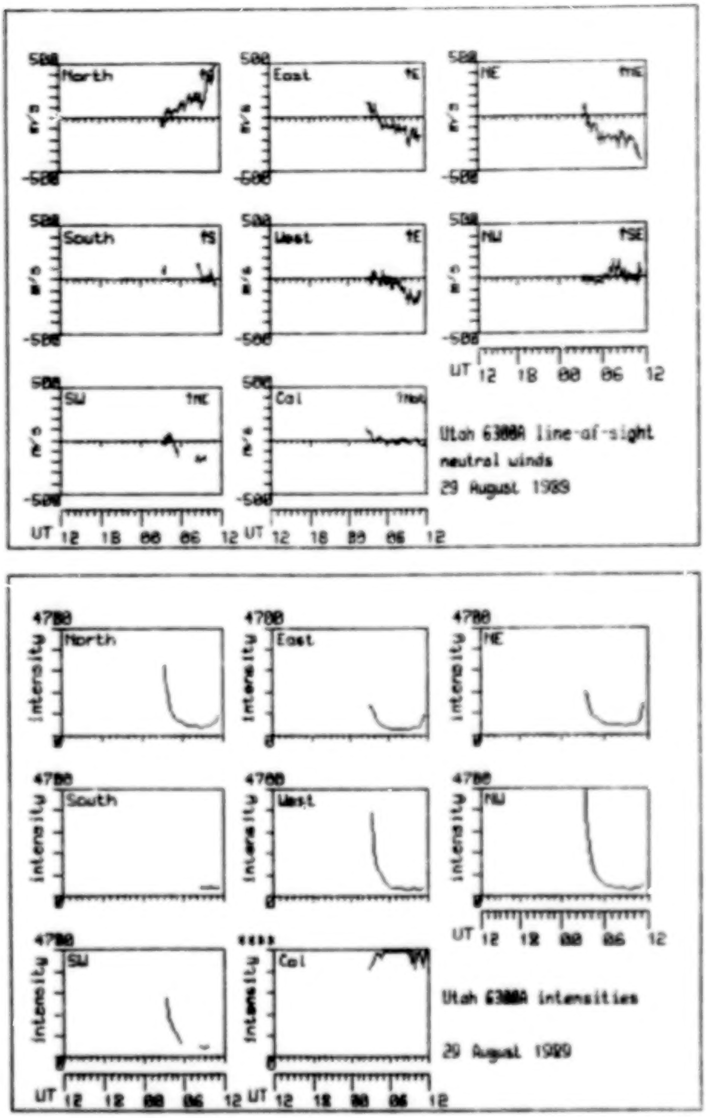


Figure 17. Initial Thermospheric wind data from the Imaging FPI at CASS, Utah, observing OI at 630 nm are presented. The data correspond to a major geomagnetic disturbance on August 29/30 1989.

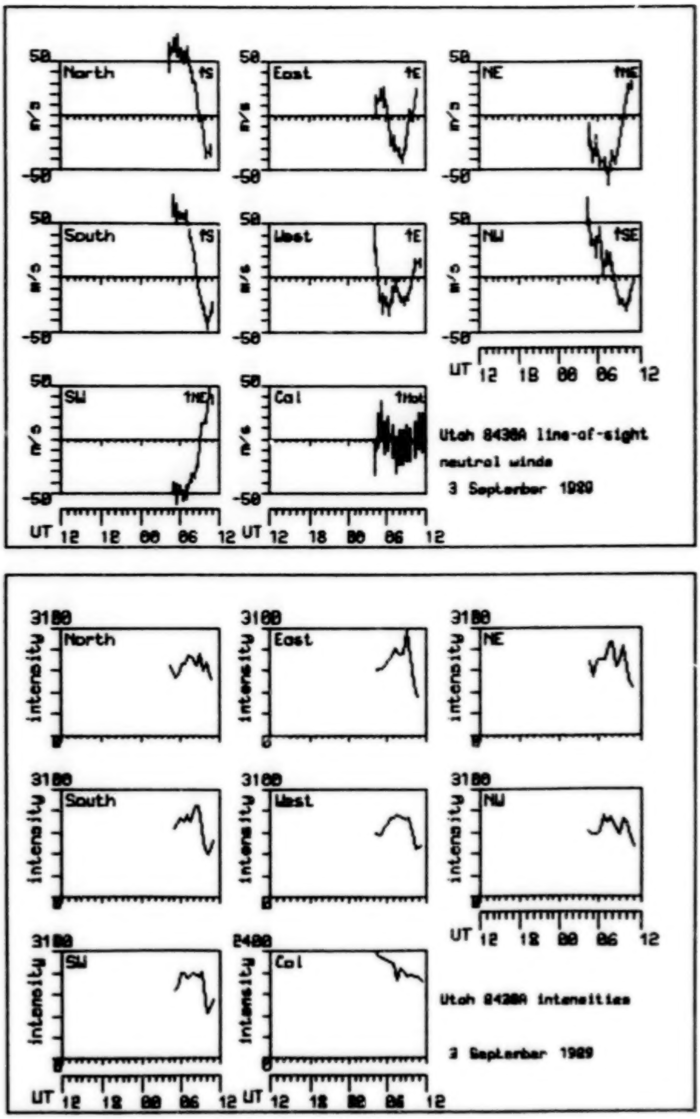


Figure 18. Initial mesospheric wind data from the Imaging FPI at Bear Lake, Utah observing OH at 843/845 nm are presented (September 3/4, 1989). There is a relatively regular 'tidal' behaviour of these winds from night to night, although wave-like features are also observed.

instrument on a routine basis. It has still, however, proven impractical to fully analyse the entire data set available. The full data sets are normally given an initial analysis, and are then archived. The data from the standard FPI is then usually used to determine a sub-set of those nights during which the more interesting thermospheric / auroral events may have occurred. With an initial selection, the full data analysis, which is a heavy user of time, computing effort and graphical hard-copy, can then be applied with rather greater efficiency. This instrument is due for a significant upgrade in the 1989 / 1990 winter: all-sky optics, detector and spectral calibration system.

Some of the data from the DIS were displayed earlier. In Figures 19 and 20, data from two disturbed nights are shown (Batten et al., 1990). Figure 19 shows data from the period 15:30 to 16:30 UT on the afternoon of December 21, 1987. There is a very large geomagnetic disturbance in progress, causing the bright OI 630 nm emissions, and the fast (500 ms^{-1}) westward winds. There are two distinct maxima in the westward wind speeds, the first near 16 UT, and the second around 16:40 UT. In Figure 20, a slightly different presentation is used to display a summary of the winds along the 20 degree east meridian, for the entire night of Dec 16 / 17 (UT) 1987. Many of the classical features of the thermospheric wind response on a night of moderate auroral and geomagnetic activity can be observed: westward winds in the dusk period, strong auroral disturbances around 16-18 UT, a rotation of the wind vector to southward and southeastward between 18 and 01 UT, and a general reduction of wind speed toward the dawn period. Just before dawn, as the auroral oval moves well poleward of Kiruna, a combination of local time effect and the general reduction of geomagnetic activity (the K index is plotted above the corresponding wind data), the winds turn anti-sunward, i.e. toward the northwest.

We are also in the process of developing a second DIS unit, which is intended to be mobile: to be used at various locations on a campaign basis. This system will probably use a Gallium Arsenide IPD, as available from ITT. As the DIS is applied to spectral targets other than OI 630 nm, we can expect that some very interesting new results will be obtained.

Acknowledgements:

The authors wish to recognise the contributions of Keith Smith and Jim Percival, in particular, to the development of the Fabry-Perot Interferometers and the Doppler Imaging System, and to Ake Steen, in particular, for general on-site support, and for maintaining the regular operation of the various instruments at IRF, Kiruna. The late Mr. Harenkowski provided continuous support for the instrument at Kilpiskarvi, a role recently continued by his wife. Dr. Risto Pellinen of Met. Inst. Helsinki, Dr. Kjell Henricksen of Tromso University, Dr. Robert Sica and Dr. Vince Wickwar of CASS, Utah State University provided the great deal of support required to get the optical facility set up at the CASS and recently the Hardware Ranch / Bear Lake Site in Utah. The instruments were all fabricated at UCL. With the exception of the Saskatchewan and Halley Bay instruments, their fabrication and the UCL aspects of their operation and the data analysis etc have been partly funded through grants from the UK Science and Engineering Research Council. The contributions of IRF, Kiruna, Tromso University (Svalbard) and CASS (Utah State University) to the maintenance of the facilities and routine operation of the respective instruments is gratefully acknowledged.

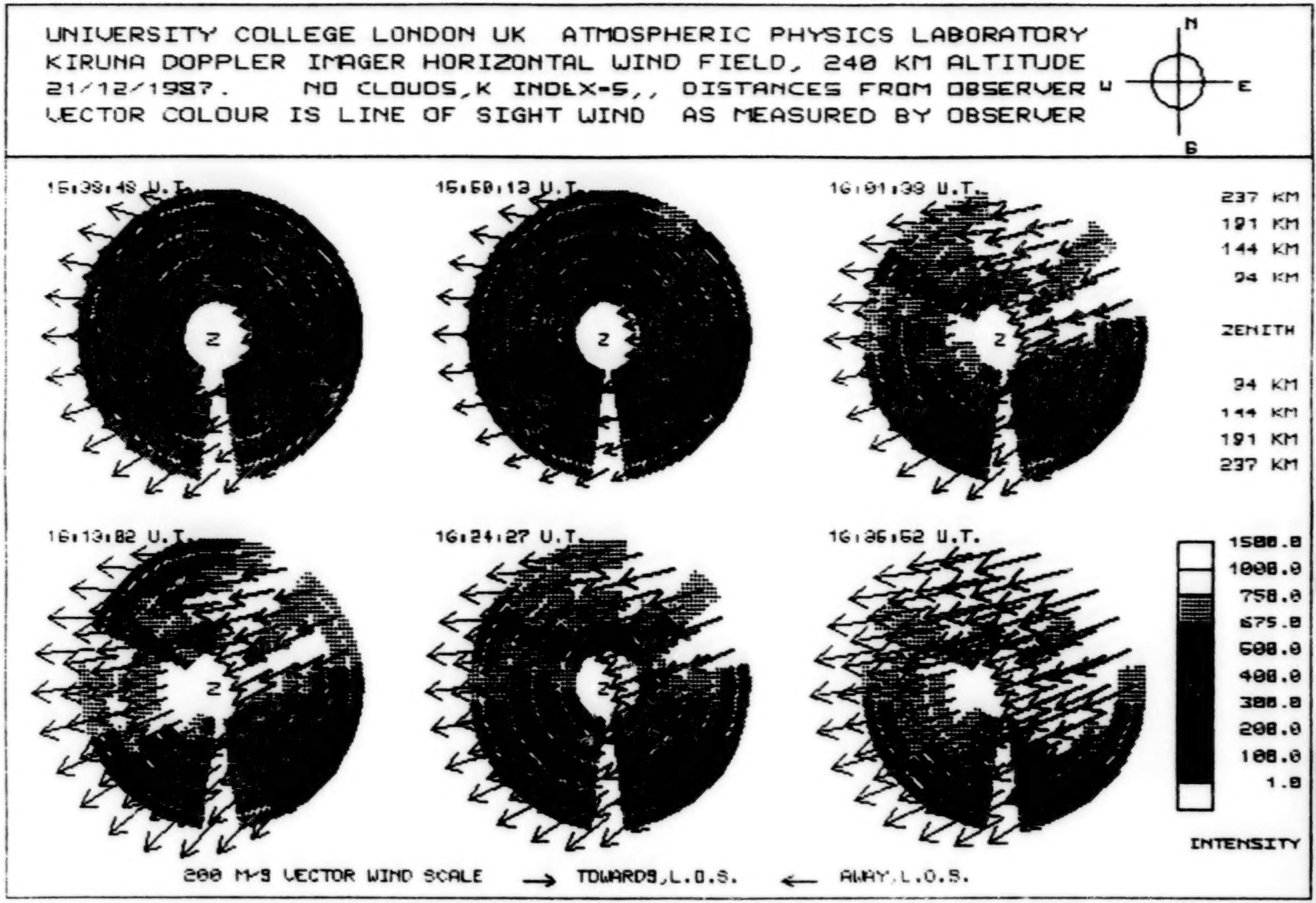


Figure 19. Data from the DIS at Kiruna during the disturbed night December 21, 1987.

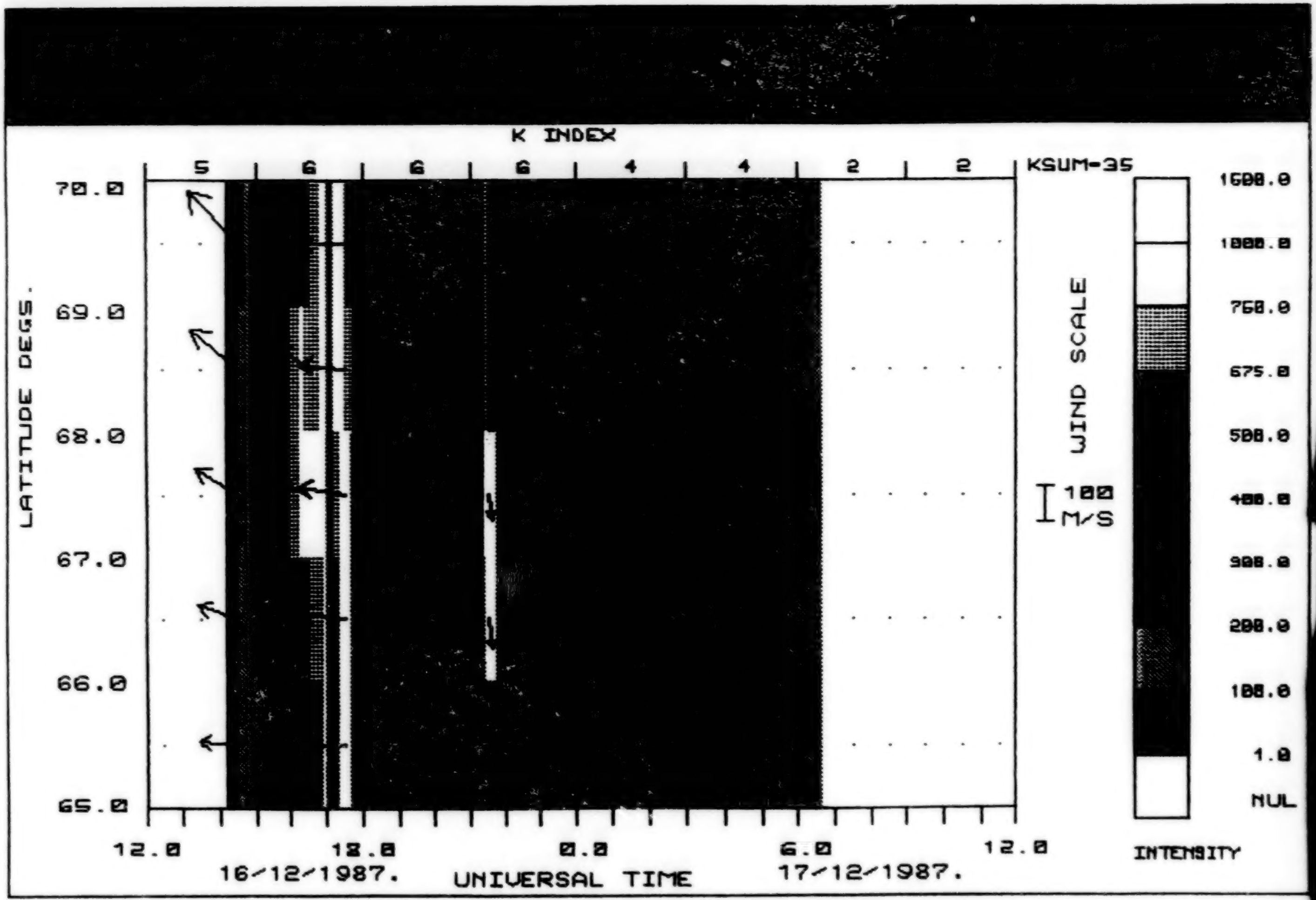


Figure 20. Data from the DIS at Kiruna during the disturbed night December 16, 1987.

220

References

Abreu V. J., G. A. Schmitt, P. B. Hays, J. W. Meriwether, C. A. Tepley and L. L. Cogger (1983). "Atmospheric scattering effects on ground-based measurements of thermospheric winds". Planet. Space Sci., 31, 303-310.

Armstrong E. B. (1968). "Variations in the width of the OIλ5577 line night airglow". Planet. Space Sci., 16, 211-229.

Aruliah A., D. Rees and T. J. Fuller-Rowell (1990). "The Combined Effect of Solar and Geomagnetic Activity on High Latitude Thermospheric Neutral Winds". Paper presented at the IAGA meeting in Exeter, July 1989, to be published in a special issue of Jn. Atmos. Terr. Phys.

Aruliah A. (1990), "The Synoptic Variety of the Thermosphere and Mesosphere Winds observed using a Fabry-Perot Interferometer". Ph. D. Thesis, University of London.

Batten S. M., D. Rees, D. Wade and A. Steen (1988). "Observations of thermospheric neutral winds by the UCL Doppler Imaging System at Kiruna in northern Scandinavia". J. Atmos. Terr. Phys. Vol 50, No 10/11, p861-888.

Batten S. M., and D. Rees (1990). "Thermospheric Winds in the Auroral Oval: Observations of small scale structures and rapid fluctuations by a Doppler Imaging System". J. Atm. Terr. Phys. (in press).

Burch S. F., S. F. Gull and J. Skilling (1983). "Image restoration by a powerful maximum entropy method". Computer Vision, Graphics, and Image Processing 23, p113-128.

Fuller Rowell T. J., D. Rees, S. Quegan, R. J. Moffett and G. J. Bailey (1987). "Interactions between Neutral Thermospheric Composition and the Polar Ionosphere using a Coupled Ionosphere-Thermosphere Model". J. Geophys. Res. 92, p7744-48.

Gull S. F. and G. J. Daniell (1978). "Image reconstruction from incomplete and noisy data". Nature, 272, p686-690.

Gull S. F. and J. Skilling (1984). "Maximum entropy method in image processing". IEE Proc, Vol 131, p646-659.

Hernandez G. (1986). Fabry-Perot interferometers. Cambridge University Press.

Kaskiewicz, P., (1978). "Interferometric Observations of Airglow Emissions from Space Vehicles", Phd. Thesis, University of London.

Killeen T. L., P. B. Hays, B. C. Kennedy and D. Rees (1982). "Stable and rugged etalon for the Dynamics Explorer Fabry-Perot Interferometer: II. Performance". Appl. Opt., 21, 3903-3912.

Killeen T. L., P. B. Hays, N. W. Spencer and L. E. Wharton (1982). "Neutral winds in the thermosphere as measured by Dynamics Explorer". Geophys. Res. Lett. 9, 957-960.

Killeen T. L. and P. B. Hays (1984). "Doppler line profile analysis for a multichannel Fabry-Perot interferometer", Appl. Opt., 23, 612-620.

Lloyd N. D. (1985). "Measurement of Neutral Thermospheric Winds using Ground-Based Fabry-Perot Interferometers", Ph. D. Thesis, University of London.

Lloyd N. D., D. J. McEwen, G. S. Sofko and D. Rees (1989). "Observations of 557.7 nm neutral winds from October 1986 to June 1987 using a ground-based Fabry-Perot interferometer at Saskatoon, Canada (52^{O}N, 107^{O}W)". J. Geophys. Res., in press.

McWhirter I., D. Rees and A. H. Greenaway (1982). "Miniature imaging photon detectors. III. An assessment of the performance of the resistive anode IPD". J. Phys. E: Sci. Instrum., 15, p145-150.

Meriwether J. W., Jr., P. Shih, T. L. Killeen, V. B. Wickwar and R. G. Roble (1984). "Nighttime thermospheric winds over Sondre Stromfjord, Greenland". Geophys. Res. Lett., 9, 931-934.

Meriwether J. W., Jr., J. W. Moody, M. A. Biondi and R. G. Roble (1986). "Optical interferometric measurements of winds at Arequipa, Peru". J. Geophys. Res., 91, 5557-5566.

Meriwether J. W., Jr. (1989). "Ground-based Fabry-Perot Interferometers": This Issue.

Rees D. (1979). "Remote Sensing of Middle Atmosphere Winds from Balloon Platforms". (COSPAR) Scientific Ballooning. Ed. W. Reidler, Pergammon Press, 207-210.

Rees D., I. McWhirter, P. A. Rounce, F. E. Barlow and S. J. Kellock (1980). "Miniature imaging photon detectors". J. Phys. E: Sci. Instrum., 13, 763-770.

Rees D., I. McWhirter, P. A. Rounce and F. E. Barlow (1981). "Miniature imaging photon detectors II. Devices with transparent photocathodes". J. Phys. E: Sci. Instrum., 14, 229-233.

Rees D., T. J. Fuller-Rowell, A. Lyons, T. L. Killeen and P. B. Hays (1982). "Stable and rugged etalon for the Dynamics Explorer Fabry-Perot Interferometer: I. Design and Construction". Applied Optics 21, 3896-3902.

Rees D. and A. H. Greenaway (1983). "The Doppler Imaging System: An optical device for measuring vector winds. I. General Principles". Applied Optics 22, 1078-1083.

Rees D., T. J. Fuller-Rowell, R. Gordon, T. L. Killeen, P. B. Hays, L. E. Wharton and N. W. Spencer (1983). "A Comparison of Wind Observations of the Upper Thermosphere from the Dynamics Explorer Satellite with the Predictions of a Global Time-Dependent Model". Planet. Space Sci. 31, 1299-1314.

Rees D., A. H. Greenaway, R. Gordon, I. McWhirter, P. J. Charleton and A. Steen (1984). "The Doppler Imaging System : Initial observations of the auroral thermosphere". Planet. Space Sci. 32, p273-285.

Rees D., R. W. Smith, P. J. Charleton, F. G. McCormac, N. D. Lloyd and A. Steen (1984a). "The generation of vertical thermospheric winds and gravity waves at auroral latitudes, I, Observations of thermospheric winds". Planet. Space Sci. 32, 667-684.

Rees D., M. F. Smith and R. Gordon (1984b). "The Generation of Vertical Thermospheric Winds and Gravity Waves at Auroral Latitudes, II, Theory and Numerical Modelling of Vertical Winds". Planet. Space Sci. 32, 685-705.

Rees D., N. D. Lloyd, P. J. Charleton, M. Carlson, J. Murdin and I. Haggstrom, (1984c). "Comparison of Plasma Flow and Thermospheric Circulation over Northern Scandinavia using EISCAT and a Fabry-Perot Interferometer". J. Atmos. Terr. Phys. 46, 545-564.

Rees D., R. Gordon, T. J. Fuller-Rowell, M. F. Smith, G. R. Carignan, T. L. Killeen, P. B. Hays and N. W. Spencer (1985). "The Composition, Structure, Temperature and Dynamics of the Upper Atmosphere in the Polar Regions during October to December 1981". Planet. Space Sci. 33, 617-666.

Rees D., T. J. Fuller-Rowell, R. Gordon, M. F. Smith, J. P. Heppner, N. C. Maynard, N. W. Spencer, L. Wharton, P. B. Hays and T. L. Killeen (1986). "A Theoretical and Empirical Study of the Response of the High Latitude Thermosphere to the Sense of the 'Y' Component of the Interplanetary Magnetic Field". Planet. Space Sci. 34, 1-40.

Rees D., N. D. Lloyd, T. J. Fuller-Rowell and A. Steen (1987). "Variations of thermospheric winds in northern Scandinavia between 1980 and 1986: a study of geomagnetic activity effects during the last solar cycle". Surv. Geophys. 9, 197-214.

Rees D., V. B. Wickwar and R. J. Sica (1989). "First OH (λ = 843 nm) Observations from the Bear Lake Site of the Hardware Ranch Observatory using a GaAs Imaging Photon Detector, Abstract for fall AGU meeting, 1989". EOS, Trans. Am. Geophys. Union.

Sibisi S. (1983). "Two-dimensional reconstructions from one-dimensional data by maximum entropy". Nature, Vol. 301, p134-137.

Skilling J. and R. K. Bryan (1984). "Maximum entropy image reconstruction: general algorithm". Mon. Not. R. Astr. Soc. 211, p111-124.

Smith R. W. and P. J. Sweeney (1980). "Winds in the thermosphere of the northern polar cap". Nature, 284, p437-438.

Steele D. P. and D. J. McEwen (1988). "Electron auroral excitation efficiencies and intensity ratios". Submitted to J. Geophys. Res.

Stewart R. D., R.W. Smith, D. Rees, A. S. Rodger and J. R. Dudeney (1985). "First Measurements of Thermospheric winds in Antarctica by an Optical Ground-based Method". Nature, 317, 45-47.

Stewart R. D., A. S. Rodger and J. R. Dudeney (1988). "Thermospheric wind response to driving forces in the vicinity of the Harang discontinuity". Planet. Space Sci., 36, 225-229.

Chapter 8

SPECTROPHOTOMETRIC MEASUREMENT TECHNIQUES OF WEAK EMISSIONS IN TWILIGHT AIRGLOW

Craig A. Tepley
National Astronomy and Ionosphere Center, Arecibo Observatory,
P. O. Box 995, Arecibo, PR 00613

SUMMARY

The measurement of weak spectral lines during twilight requires a careful analysis treatment for the removal of scattered solar Fraunhofer structure. We first describe the process of resonant scattering and discuss the geometry encountered during a twilight experiment, including the effects of the transmission of the lower atmosphere to the excitation radiation. This is followed by a description of an analysis procedure which stresses the need to account for an additional continuum other than the measured solar and nightglow spectra. Some observations of the temporal and spectral behavior of this extra continuum are also given.

1. INTRODUCTION

Twilight is by definition that period just after sunset and before sunrise when a portion of the atmosphere is illuminated by the sun's rays from below, contrary to daylight when the rays enter the atmosphere from above. As the solar zenith, or depression angle increases, the lower atmosphere becomes progressively darker as the rays from the sun continue to rise in the atmosphere. This condition permits a variety of structure and composition studies of many minor constituents of the atmosphere by measuring their scattered resonant or fluorescent radiation from the sun. As the strong scattered light from the solar continuum decays, observations from the ground of the weaker resonance lines of many species are feasible, compared to what would otherwise be an impossible task in daylight. In the absence of solar illumination at night the measurement of species through their resonant excitation is also impossible unless an artificial source such as a lidar is employed.

Although somewhat limited in the range of the species that can be observed, ground-based optical observations using spectrophotometric techniques during twilight offer an inexpensive means to study the minor constituents of the middle atmosphere compared with rocket observations. For a particular region of the atmosphere, for example between 50 to 150 km altitude, twilight may last for several hours in the polar regions but only 15 or 20 minutes near the equator, depending on the season. Even during short twilights if sufficiently fast data sampling techniques are employed, such as programmable wavelength scanning spectrometers [Meriwether, 1979], and with knowledge or good estimates of the transmission of the atmosphere to the excitation energy, the inversion of the scattered emission profile will yield reasonable estimates of the altitude distribution of the species under study. Nevertheless, except for only a few cases, this powerful technique has not been fully exploited since the pioneering work that was accomplished in the 60's.

It is not the purpose of this report to provide a complete and exhausted review of all twilight measurements that have been made to date. We regret any omissions that would otherwise deserve proper reference. We will try to outline, however, some of those accomplishments with particular emphasis on investigations of the middle atmosphere, where future efforts would play important roles in understanding the chemistry and dynamics of that region, and which are an important part of the WITS program with regard to composition studies. Next, we offer a brief description of the observational method used to acquire spectral information, first through a synopsis of the scattering process followed by the experimental geometry encountered in twilight. The paper concludes with a detailed discussion of the data reduction techniques needed for the analysis of the spectra obtained, and of the relevant errors involved. Emphasis is placed on the need to account for a quasi-continuum in the analysis which may be the result of a lingering fluorescence of aerosols from the lower atmosphere during the twilight periods, and which can seriously affect the ability to obtain accurate magnitudes of weak line emissions.

2. OBSERVATIONS

Hunten [1967] gave a very thorough account of twilight airglow studies that occurred before the late 60's. He discussed several of the nightglow emissions that were observed to have twilight enhancements, and summarized those which can only be measured during twilight periods. Some notable examples of these include the thermospheric emissions of the first negative band system of N_2^+ [Lytle and Hunten. 1962; Broadfoot and Hunten, 1966], the metastable states of $O(^1D)$ and $N(^2D)$ [Barbier, 1959; Noxon, 1964; Dufay, 1953], and the infrared $He(^3P)$ line at 1.08 μm [Fedorova, 1961]. Emissions from the lower thermosphere which can be observed in twilight include the atmospheric and infrared atmospheric bands of O_2 [Vallance Jones and Harrison, 1958] and the resonant lines of various metallic atoms and ions, in particular those from the alkali metals [Broadfoot, 1967; Chamberlain, 1961; Gadsden, 1969; see also Hunten, 1967].

In the two decades following the earlier work briefly outlined above, there have been only a few ground-based twilight measurements of emissions from the upper ionosphere and thermosphere. These are the observations of $O^+(^2P)$ [Meriwether et al., 1978], the emissions from the metastable ^1S and ^1D states of N^+ [Torr et al., 1981; 1983], and more recently of N_2^+, an example of which is shown in Figure 1 [Torr et al., 1989]. Twilight enhancements have been observed from balloons and from the ground for the (0-1) component of both the molecular oxygen bands mentioned above [Evans et al., 1970; Pick et al., 1971; Noxon, 1975; Lowe et al., 1984]. The enhancement arises from both fluorescent scattering of sunlight and from photochemistry, the latter of which can provide estimates of the ozone concentration below about 100 km [Evans and Llewellyn, 1972; Noxon, 1975].

The presence of metallic species in the lower ionosphere from about 70 to 125 km can provide important tracers of the dynamics and chemistry of that region. However, very few attempts have been made during the last twenty years with ground-based observations using resonant scattering from the sun in twilight in order to investigate

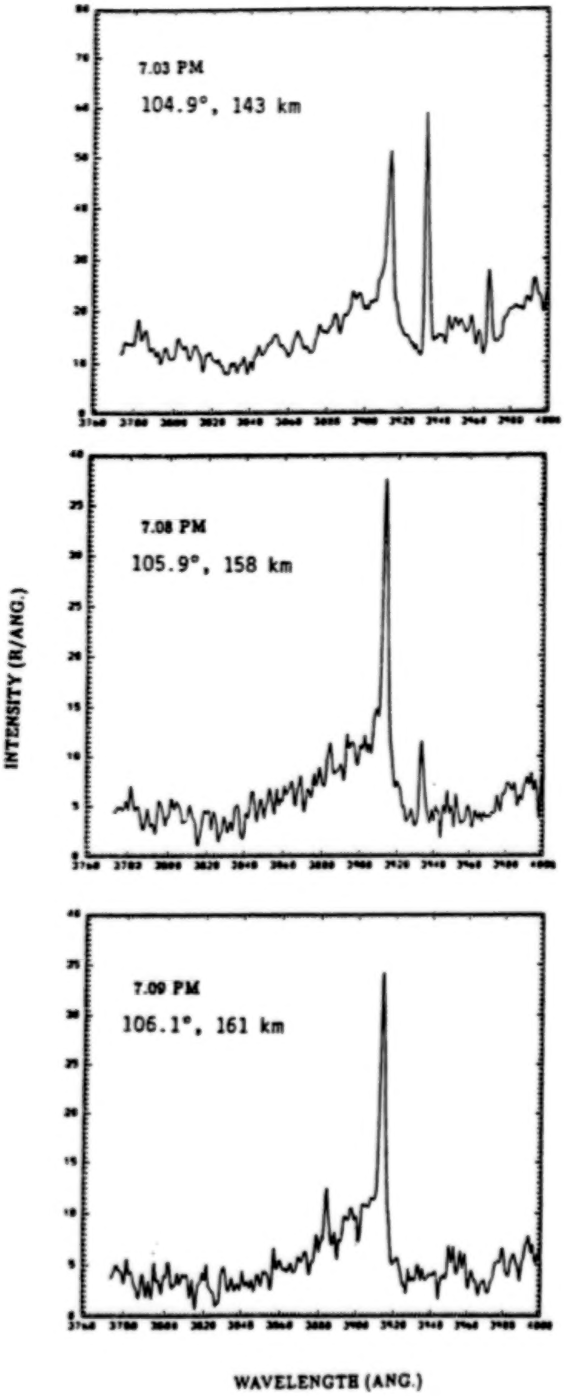


Figure 1 - Three twilight spectra measured on the evening of 22 December 1987 by Torr et al. [1989] showing the rotational distribution of the (0-0) and (1-1) bands of the N_2^+ 1st Negative system at 3914Å and 3884Å, respectively, and the appearance of the H and K lines of Ca^+ at 3968Å and 3934Å. (Figure courtesy of D.G. Torr).

the behavior of these metals. There have been a few polar and equatorial observations of Li and Na [Henriksen et al., 1980; Kirchhoff and Clemesha, 1973], and those of Fe and Ca^+ from low and mid-latitudes sites [Broadfoot and Johansen, 1976; Tepley et al., 1981a,b; Torr et al., 1989]. Observations of a metal in both its neutral and ionic form will help unravel the chemical distribution of these species. The ionic states of the more abundant metals as measured by mass spectrometers using rockets, such as Fe^+, Mg^+, and Ni^+, have resonance lines in the ultraviolet. Unfortunately, the lesser abundant calcium is the only metal which can be observed from the ground as both Ca and Ca^+. Although somewhat rare in occurrence, one such twilight detection for example, of the resonant emission from neutral calcium at 4227Å is shown in Figure 2. The progressive decay of the strength of this line with increasing solar depression angle can be compared with the average variations of similar observations of Ca^+ made from the same location [see figure 2 of Tepley et al., 1981b]. Ratios of the derived concentrations, that is $[Ca^+] / [Ca]$, (or column abundances) range from 0.7 near 91 km to 3.0 at 98 km altitude, which are consistent with chemical models of these metals in this altitude region [Brown, 1973].

Spectroscopic measurements of weak emissions in twilight with ground-based instruments is a powerful tool to study the composition of the atmosphere. Future observations can still take advantage of this useful and relatively inexpensive technique. The method, however, has been eclipsed in the last ten years by more powerful lidar remote sensing of the metals, in particular in detailed investigations of the alkalis [Megie et al., 1978; Clemesha, 1984; Gardner and Shelton, 1985], but also more recently with measurements of Fe, Ca, and Ca^+ [Granier et al., 1989a,b]. Lidar studies are not necessarily restricted to nighttime observations but can also be made during the day [Granier and Megie, 1982; Kwon et al., 1987]. The technique lends itself, not only to the composition studies mentioned above, but to direct measurements of the temperature of the mesopause region [Fricke and von Zahn, 1985; von Zahn and Neuber, 1987].

3. SPECTROSCOPIC BACKGROUND

Here we outline some of the basic spectroscopic theories encountered. Much of what we describe with regard to resonant scattering is detailed in two excellent texts by J.W. Chamberlain, *Physics of the Aurora and Airglow* and *Theory of Planetary Atmospheres*. The former text is almost 30 years old but is still a very useful standard reference for all aspects of airglow and auroral studies. The notation used in the rest of this section is similar, and in many ways identical, to that found in Chamberlain's texts.

The process of resonant scattering in an optically thin atmosphere will be treated first. This is the simplest form of scattering ignoring such effects as multiple scattering or the deactivation of the scatterers before emission occurs, otherwise known as quenching. When the atmosphere is considered optically thick, that is, the intensity of the emission is strong while the concentration of the emitters is large, the equations are still tractable. However, for weaker lines, such as metallic radiations in twilight for example, this is not the case (that is, optically thin conditions prevail) and we will show that the simple theory is valid. In the following section the additional problems

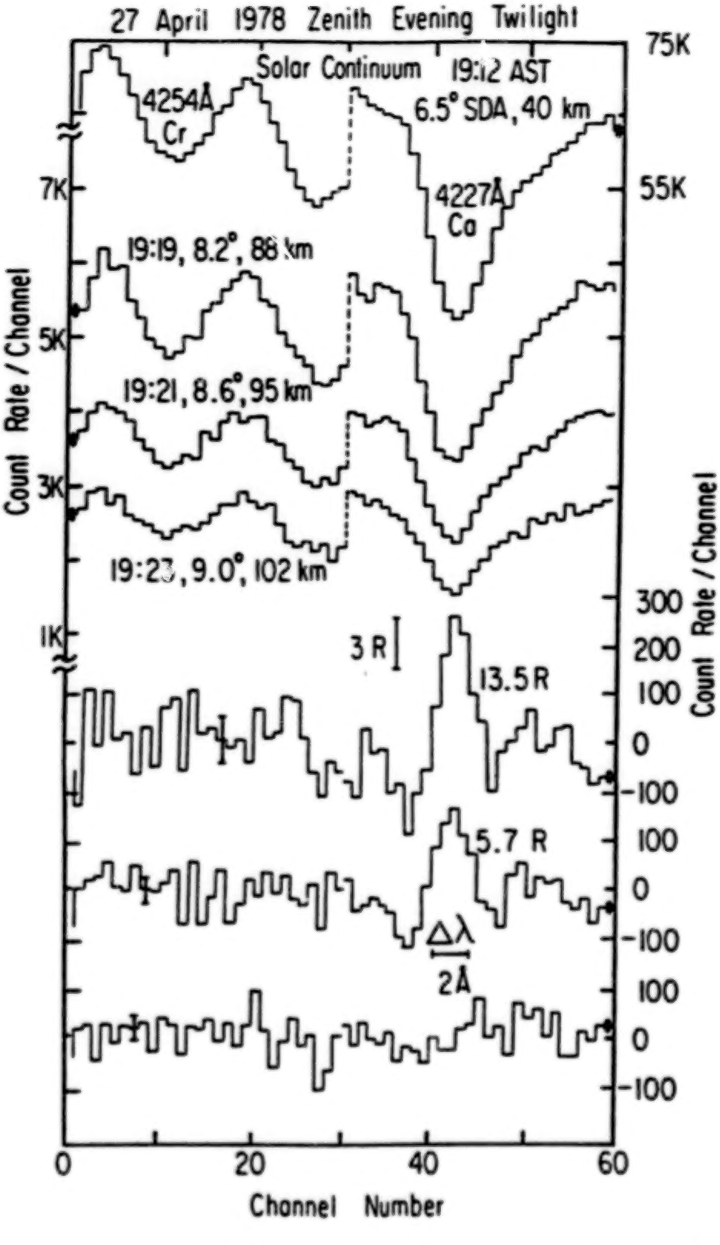


Figure 2 - Detection of the Ca(4227Å) resonance line in the evening twilight of 27 April 1978. The top spectrum is the scattered solar continuum measured in early twilight showing the Ca and Cr(4254Å) Fraunhofer absorption features. A programmable wavelength scan was used to obtain these spectra over two short ranges near the resonance lines of these metals [see Meriwether, 1979]. The next three spectra shown below the solar continuum were measured from middle twilight between about 8° and 9° solar depression. The bottom three curves are the results of applying the data reduction technique described in Section 5 to the middle twilight spectra, showing the detection of Ca and the absence of the Cr resonance.

encountered when dealing with resonant emissions are also discussed. These include the effects of atmospheric absorption and secondary scattering of both the incident and radiated energies, generally called extinction, and the atmospheric screening and transmission of the radiation. We can view these extra effects as an extension to the simply theory.

3.1 Resonant scattering

Before we describe the governing equations characteristic of resonant scattered emissions, we first outline some basic concepts of airglow photometry. In photometric studies it is usually common to express the surface brightness of some distant source in units of erg cm^{-2} s^{-1} str^{-1}. In airglow and auroral work the energy unit is replaced with the number of photons for convenience. If the source at a distance r from the observer is characterized by a volume emission rate given by E(r) photons cm^{-3} s^{-1}, then integrating with respect to the column distance yields its surface brightness, or

$$4\pi I = \int_0^\infty E(r)\, dr \tag{3.1}$$

where $4\pi I$ has the units photons cm^{-2} s^{-1}. This leads to the concept of the Rayleigh. If I has the units of 10^6 photons cm^{-2} s^{-1} str^{-1}, then $4\pi I$ is equal to one rayleigh (R) emission.

Resonance scattering of solar radiation is a convenient way of identifying certain atoms and molecules. The resonance lines of these species can be thought of as their fingerprint, and are usually identified by the transitions from the lowest lying excited energy level to the ground state of the species. Resonance emission is distinguished from fluorescence in that the absorbed solar energy is reradiated at the same frequency. Fluorescence is energy reradiated by a route within the energy levels of an atom or molecule which is different from the absorption path. Consider a situation where solar radiation is uniformally scattered by a collection of atoms. If there is no deactivation of the resonant energy through collisions before reradiation, then for a solar flux per unit area at a frequency ν penetrating to an altitude z is given by $\pi F_\nu(z)$, the energy absorbed is $\pi F_\nu(z) n(z) \alpha_\nu$, where n(z) is the number density of atoms at height z which are in their ground state and α_ν is the absorption coefficient. The absorption coefficient for a particular state of an atom is related to the transition probability A and the statistical weight $\bar{\omega}$ of that state. Summing over all frequencies we have

$$\int \alpha_\nu(0,1)\, d\nu = \frac{\pi e^2}{m_e c}\ f(0,1) = \frac{c^2}{8\pi\nu^2}\ \frac{\bar{\omega}_1}{\bar{\omega}_0}\ A(1,0) \tag{3.2}$$

where f is the oscillator strength (sometimes called the f-value) for the state and the subscripts 1 and 0 denote the upper and ground states, respectively. We have used 0,1 instead of a more general i,j notation to signify a resonance, where radiation is resonantly scattered in the transition 1→0. The other symbols have their usual meanings. Finally, the intensity of the scattered column emission (in rayleighs) along the line of sight is

$$4\pi I = \int_{z_o}^{\infty} n(z)\ [\pi F_\nu(z)]\ \frac{\pi e^2}{m_e c}\ f(0,1)\ dz \qquad (3.3)$$

This is the simplest form of the scattering equation.

To allow for deactivation as mentioned above, we can multiply the term under the integral in equation (3.3) by an effective albedo term which accounts for the ratio of transition probabilities, or

$$\bar{\omega}(1,0) = \frac{A(1,0)}{A(1,0) + \eta(1,0)\ [X]} \qquad (3.4)$$

Here, η is the rate coefficient for collisional deactivation of the (0,1) state, or simply the collision frequency, and [X] is the concentration of all non-(0,1) species, which is also a function of altitude. For unattenuated solar radiation, it is common practice to collect most of the terms of equation (3.3) into a constant for the particular state of the atom. Representing this constant as the so-called "g-factor" we have

$$4\pi I = g(0,1)\ n_{tot} \qquad (3.5)$$

where n_{tot} represents the column density of the scatterers and g(0,1) is defined as

$$g(0,1) \triangleq \pi F_\nu(\infty)\ \frac{\pi e^2}{m_e c}\ f(0,1) \qquad (3.6)$$

Thus, if we can measure the intensity of a particular resonance emission, we can find the integrated column density of the scatterers through their g-factors. The g-factors for various resonance lines of atoms and molecules have been computed and tabulated by Chamberlain [1961], Hunten [1967], and Gadsden [1969]. We repeat a table of the values (Table I) for some of the weaker features observed. It must be remembered that the value of g is found for unattenuated solar radiation. Modifications to the theory must be made when the radiation is partially absorbed or scattered by other atmospheric species before reaching the scattering volume [Swider, 1984]. These extra effects are illustrated in the next section.

Finally, the above treatment ignores two factors which may be important for certain types of radiations. The first is the optically thick atmosphere which effects the incident radiation. This adds an additional albedo term to the solar flux which splits the energy into absorbed and scattered components. It is unimportant, for example, for twilight metallic resonance observations as we show in an example below, but it could conceivably be important for corresponding dayglow measurements. The second effect is the polarization or anisotropy of the scattered radiation. This can be accounted for by adding a phase function to the scattered intensity equations given above as (3.3) and (3.5), where this phase function delineates the angular distribution of the scattered radiation. These extra effects have been detailed by Chamberlain [1978] and will not be repeated here.

Table I
Scattering Efficiencies for the Observed Species in Twilight

Species	Line(Å)	g(ph s^{-1})	Species	Line(Å)	g(ph s^{-1})
Fe	3860	0.014	Na D_2	5890	0.540
N_2^+ 1Neg	3914	0.050	Na D_1	5896	0.330
Ca^+ K	3934	0.270	Hα	6563	2.6×10^{-6}
Al	3962	0.041	Li	6708	8.7
Ca^+ H	3968	0.150	K D_1	7699	1.920
Mn	4031	0.037	O_2(0-1) Atm	8645	5.0×10^{-10}
Ca	4227	0.310	He(^{3}P)	10830	16.8
Cr	4254	0.037	O_2(0-0) IR Atm	12683	9.4×10^{-11}
N(^{2}D)	5200	7.5×10^{-11}	O_2(0-1) IR Atm	15803	1.2×10^{-11}

We now illustrate an example for an optically thin atmosphere. The absorption resonance scattering cross section is described by

$$\alpha_\nu = \frac{\pi e^2}{m_e c} f \frac{c}{V\nu\sqrt{\pi}} \exp\left[-\frac{c^2}{V^2}\frac{(\nu-\nu_o)^2}{\nu_o^2}\right] \text{cm}^2 \tag{3.7}$$

where the integrated absorption coefficient per atom per unit f-value is $\pi e^2/m_e c$ (=0.02647 $cm^2\,s^{-1}$, viz. equation 3.2), f is the oscillator strength, V is the thermal velocity ($\sqrt{2\kappa T/M}$) of the atoms of mass M, and ν_o is the center frequency [Chamberlain, 1961]. The other terms in the equation have their usual definitions. We next determine if the atmosphere is optically "thick" to an emission. Using the K line of the calcium ion as an example (at $\lambda = \lambda_o = 3934$Å, f=0.69), and at T=200 K, the absorption coefficient is equal to 1.4×10^{-11} cm^2. This translates to a column abundance of about 7×10^{10} cm^{-2}. Using (3.5) and the appropriate g-factor from Table I, we see that for unit optical depth, the Ca^+ intensity must approach 19 kR. Measurements of the calcium ion intensity usually never exceeded 100 R, [Broadfoot, 1967; Tepley et al., 1981b], so, we can conclude that for this emission at least, the atmosphere is optically thin and as a result, the above equations for resonance scattering are valid.

Remember, the application of the g-factors in this way of course assumes that the excitation energy is only from direct solar flux. In actuality, this approach fails to account for the albedo of the atmosphere and reflections from the surface of the Earth, that is, backscattered radiation can excite resonances. For dayglow observations, the radiation due to the albedo is important as shown by McElroy and Hunten [1966], which leads to an increased excitation of the resonance energy resulting in an effective increase in the value of the g-factor. This would lower the determination of the column density of the scatterers. In twilight, however, when the excitation radiation originates from below the atmospheric region that is being observed, the albedo effects are minimal, being overshadowed by such things as screening of the radiation in the lower atmosphere. The effects of albedo in twilight have been estimated for sodium by Chamberlain [1956] and Donahue et al. [1956] where they found that only a few percent correction should be applied.

4. TWILIGHT GEOMETRY

An optical twilight experiment is characterized by the sun being depressed a few, to a few tens of degrees below the horizon. During this time it is possible to obtain reasonable height information of the scatterers that could not otherwise be determined from nightglow work by recording the spectrum of the species with time resolutions of the order of ½° to 1° solar depression angle. Additionally, as the sun becomes depressed, much of the unwanted background intensity decays allowing for the detection of weaker emissions that cannot be seen in daytime.

Figure 3 illustrates the geometry encountered in twilight observations. We again follow the development given by Chamberlain [1961]. There are two important concepts which can be obtained from this figure, the geometrical shadow height and the screened shadow height. The latter is the result of extinction of the excitation radiation from the sun as it passes through the troposphere and stratosphere. This extinction is due to molecular (Rayleigh) scattering, and to the presence of ozone and aerosols which are responsible for absorption and scattering of the radiation. If we restrict observations to the plane of the solar azimuth, then from the figure we have six variables of interest. These are the angles α, β, γ, χ, θ, and height z_s. Here, α is defined as the solar depression angle with respect to the observer, and χ is the zenith angle of observations. Of these six parameters, only two are known, α and χ. For example, values of α can be calculated from knowledge of the solar declination, azimuth, and observational times.

The solution to the twilight problem requires four equations. Restricting the problem to an unscreened solar ray, from the triangle containing angle β in Figure 3, we have

$$z_s = \mathrm{Re}(\sec\beta - 1) \tag{4.1}$$

where Re is the radius of the Earth. Using the law of sines for the triangle containing angle γ, we can define an equation used often in nightglow work, the van Rhijn equation,

$$\sin\theta = \frac{R_e}{R_e+z_s}\sin\chi = \cos\beta\,\sin\chi \tag{4.2}$$

We also have

$$\theta = \chi - \gamma \tag{4.3}$$

And finally, from the depression angle relationships

$$\sin\beta = \sin(\alpha-\gamma) = \sin\alpha\,\cos\gamma - \sin\gamma\,\cos\alpha \tag{4.4}$$

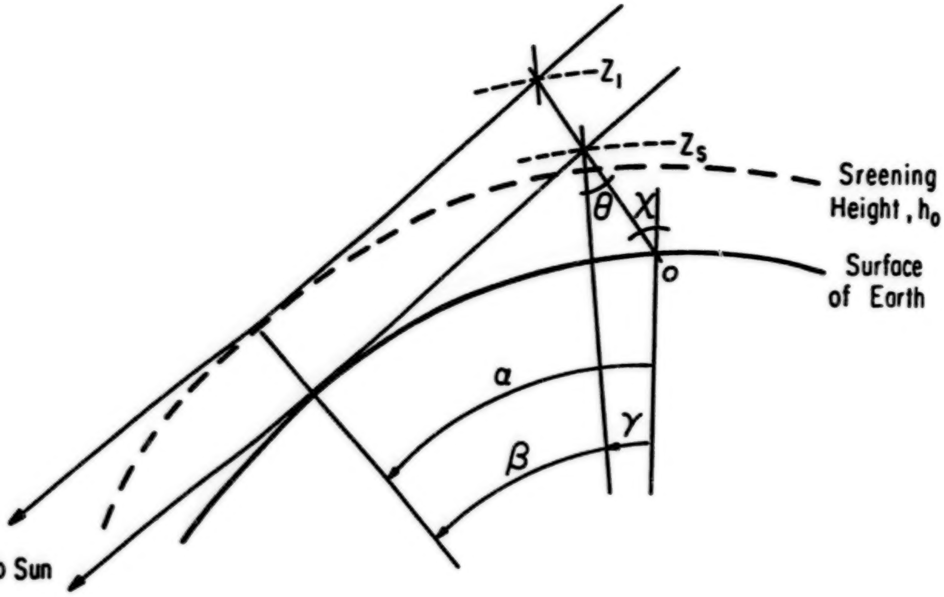


Figure 3 - The geometry encountered in twilight illustrating the screening of radiation from the sun by the lower atmosphere. The ray or line from the sun to Z_s grazes the solid earth and casts a shadow into the atmosphere at Z_s. The ray to Z_1 is the screened shadow line which "grazes" the screening height h_0. These shadow lines are not as distinct as pictured but are smeared in altitude (see text). The effects of refraction are also not shown. Observations from point O are made at a zenith angle χ; the other angles are described in the text. Generally, $\alpha \neq \beta + \gamma$ if the observations are not made in the plane of the sun's azimuth with respect to O [see Chamberlain, 1961].

When the observations are not in the plane of the solar azimuth, Chamberlain [1961] showed that equation (4.4) can be modified by including a factor, $\cos\phi$, multiplied by the last term, where ϕ is the azimuth angle between the sun and the observer. All of the other equations above remain the same. Additionally, for the simple case of zenith observations, $\alpha=\beta$ and $\theta=\gamma=\chi=0°$, only equation (4.1) remains. When the screening height h_o shown in Figure 3 is accounted for, then equation (4.1) becomes

$$z_1 = Re\ (\sec\beta - 1) + h_o \sec\beta \tag{4.5}$$

(For $h_o=0$, (4.5) reduces to (4.1) and $z_1=z_s$.) The other equations follow accordingly where, for the screened ray, γ increases while β decreases slightly, and α remains the same.

Finally, we should mention that the above analysis fails to account for the refraction of the incident solar ray. The actual ray would be bent toward the Earth by an amount

$$\Delta z \approx 127 \ \frac{N(h_o)}{N(0)} \ \frac{\sin\beta}{\cos^2\beta} \tag{4.6}$$

for $h_o << Re$ and where $N(z)$ is the total neutral density at altitude z. The factor 127 (=0.02 Re) has been measured from stellar occultations and applies for heights given in kilometers [Chamberlain, 1961]. For typical neutral concentrations, we find that refraction amounts to less than a 1% decrease in the shadow heights.

As a quick rule of thumb estimate, and neglecting proper dimensionality, we see that the height z_s in kilometers is approximately equal to the solar depression angle squared, that is, $z_s \approx \alpha^2$, where α is in degrees. If we include the atmospheric screening, then $z_1 \approx \alpha^2 + h_o$. These approximations appear to be good to 1-2% of the actual heights more rigorously calculated for the lower ionosphere.

4.1 Twilight screening and atmospheric transmission

The optical thickness of an atmosphere for a particular emission is related to the atmospheric transmission. In twilight when the shadow line is not sharply defined, then equation (3.5) must be modified by including a transmission function. Additionally, the screening height for a particular emission wavelength can be obtained from knowledge of the transmission of the lower atmosphere. The procedure has been outlined by Hunten [1962]. He used tables of molecular composition, ozone and dust concentrations, and accounted for refraction and the finite size of the sun. The resulting atmospheric transmission functions for various emissions, such as Na, Ca^+, and N_2^+, appeared somewhat Gaussian in shape having a mean value equal to the screening height and a standard deviation of one half the "smearing width" of the shadow line. No seasonal variations of the concentration of aerosols or ozone were taken into account which can alter the altitude of the screening height. In an attempt to account for the uncertainties due to variations of atmospheric transmission, Tepley et al., [1981b] used additional information of the ionospheric structure determined from simultaneous electron density measurements and optical observations of Ca^+ in twilight. By assuming the calcium ions were distributed similarly to the electrons in layers, the electron density profile was then convolved with an atmospheric transmission function with zero screening and integrated with respect to altitude to simulate a column emission rate profile for the layered emitters. A comparison of the altitude distribution of the measured and synthesized optical profiles yielded the screening height.

The procedure is as follows. Referring to Figure 4 adopted from Hunten [1962], the observed emission $4\pi I(z_1)$ is the convolution of density $n(z)$ and the transmission function $T(z-z_1)$,

$$4\pi I(z_1) = g \int_0^{\infty} n(z)T(z-z_1)\, dz \tag{4.7}$$

where g is the efficiency factor for resonant scattering mentioned previously. Equation (4.7) is a variation of (3.5) where z_1 is the screened shadow height. As shown in Figure 4, if we assume a Gaussian shape for the derivative of T with respect to z_1, then T is defined as the distribution function of the Gaussian probability density with zero mean,

$$T(z-z_1) \triangleq F(z_1) = \frac{1}{\sigma\sqrt{\pi}} \int_{-\infty}^{z_1} \exp\{-x^2/2\sigma^2\}\, dx \tag{4.8}$$

or

$$F(z_1) = \tfrac{1}{2}\,[1+\mathrm{erf}(z_1/\sigma\sqrt{2})] \tag{4.9}$$

where we have made use of the relationship between the integral of the Gaussian distribution and the standard error function which facilitates encoding into a computer program. After a suitable background density is removed, $F(z_1)$ is then convolved with a trial $n(z)$ and the result is integrated numerically throughout the region where the emitters would be distributed.

5. TWILIGHT SPECTRAL DATA ANALYSIS

In this section we present a mathematical representation for the analysis of optical spectra measured during twilight. In order to obtain accurate values of any line emissions, all continua such as the scattered solar Fraunhofer structure, must first be removed. The technique is particularly useful for its application to weak twilight emissions such as resonance scattering from metallic species. The measured radiance during any given twilight period is composed of the solar and nightglow continua, the spectral distribution of line emissions, a quasi-continuum which may be the result of some fluorescent scattering mechanism, and instrumental noise. The relative distribution of these individual components might be used to infer the amount of fluorescence scattering, possibly due to aerosols within the atmosphere. In a later section we suggest

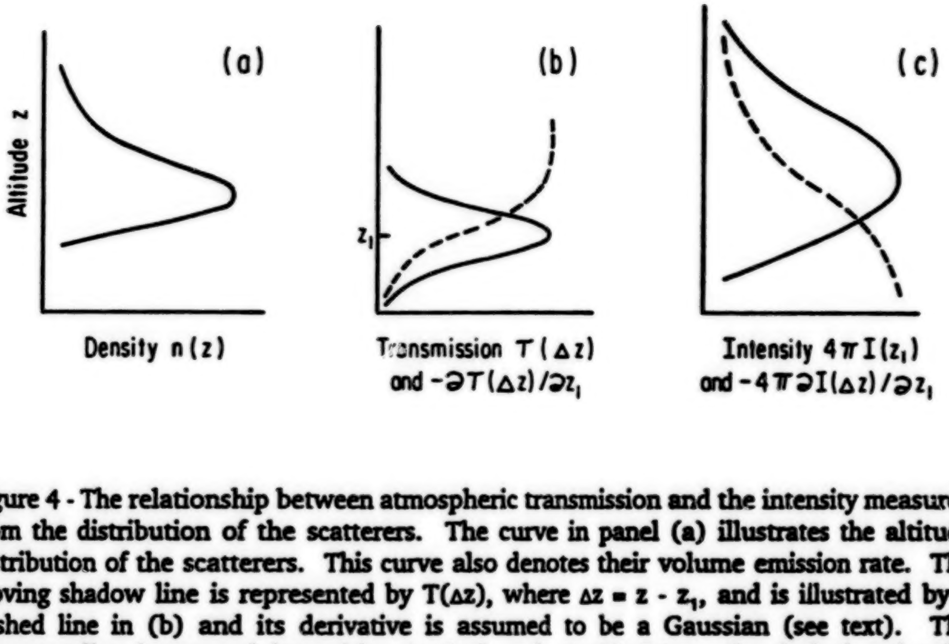

(a)
Altitude z
Density n(z)
(b)
z_1
Transmission $T(\Delta z)$ and $-\partial T(\Delta z)/\partial z_1$
(c)
Intensity $4\pi I(z_1)$ and $-4\pi\partial I(\Delta z)/\partial z_1$

Figure 4 - The relationship between atmospheric transmission and the intensity measured from the distribution of the scatterers. The curve in panel (a) illustrates the altitude distribution of the scatterers. This curve also denotes their volume emission rate. The moving shadow line is represented by $T(\Delta z)$, where $\Delta z = z - z_1$, and is illustrated by a dashed line in (b) and its derivative is assumed to be a Gaussian (see text). The corresponding integrated intensities $4\pi I(\Delta z)$ and $-4\pi\partial I(\Delta z)/\partial z_1$, the dashed and solid curves in panel (c), respectively, are the convolutions of n(z) and their respective transmission functions. If the shadow line were sharply defined, then T would be a step function and ∂T a delta function at z_1. For that case, ∂I would have the same shape as n(z) with altitude [see Hunten, 1967].

a method by which this inference might be accomplished.

We call the analysis method the ratio algorithm. Another technique used to analyze twilight spectra involves scaling the solar continuum and subtracting it from the twilight spectra. That technique (call it "differencing") has met with some experimental success [Broadfoot and Hunten, 1966; Broadfoot, 1967] but cannot adequately account for an atmospheric fluorescence component which varies in wavelength. We demonstrate how ratioing is superior to differencing in terms of minimizing the accumulation of error inherent in the data analysis procedure. The following development can also be found in Tepley et al. [1981a].

5.1 Data reduction technique

To illustrate the ratioing procedure, we define $S(\alpha_0,\lambda)$ and $F(\alpha_0,\lambda)$ to be the amplitudes of the scattered solar continuum and the fluorescence continuum, respectively, for the solar depression angle α_0. Referring to Figure 5, these two components are shown by curves (a) and (e). The amplitude of the line emission represented by the function $L(\alpha_0,\lambda)\delta(\lambda-\lambda_0)$ is small compared with the values of S and F. The wavelength variable and the line position are denoted by the quantities λ and λ_0, and $\delta(\lambda-\lambda_0)$ is the Dirac delta function.

We assume that the spectral variation for either continuum is small within the resolution of the measuring instrument, or $\Delta\lambda$. This allows us to ignore the necessity of convolving the spectral shape of $\Delta\lambda$ with the measured spectra in the development that we present below.

At α_0, the measured spectrum is written as

$$M(\alpha_0,\lambda) = S(\alpha_0,\lambda) + F(\alpha_0,\lambda) \tag{5.1}$$

At a time further into twilight at a solar depression angle α, the amplitude of the line emission L becomes a significant part of the spectrum as compared with the attenuated amplitudes of the scattered solar and fluorescence continua, defined by the scale factors, a and b, respectively. This later spectrum is represented as

$$M(\alpha,\lambda) = aS(\alpha_0,\lambda) + bF(\alpha_0,\lambda) + L(\alpha,\lambda)\delta(\lambda-\lambda_0) \tag{5.2}$$

To remove the effects of the wavelength structure of the scattered solar continuum and the fluorescent continuum from the measured spectrum, we follow the procedure of Noxon et al. [1979] as modified to treat the problem of fluorescence contamination. For each spectrum we introduced a function $C(\alpha,\lambda)$ that may vary linearly in wavelength which we use to modify the amplitude of the measured spectrum, equation (5.2), by subtraction. This resultant is then divided by the spectrum $M(\alpha_0,\lambda)$. Dropping the depression angle arguments but introducing a subscript "$_0$" to denote the spectrum measured at α_0, these ratios are then given by

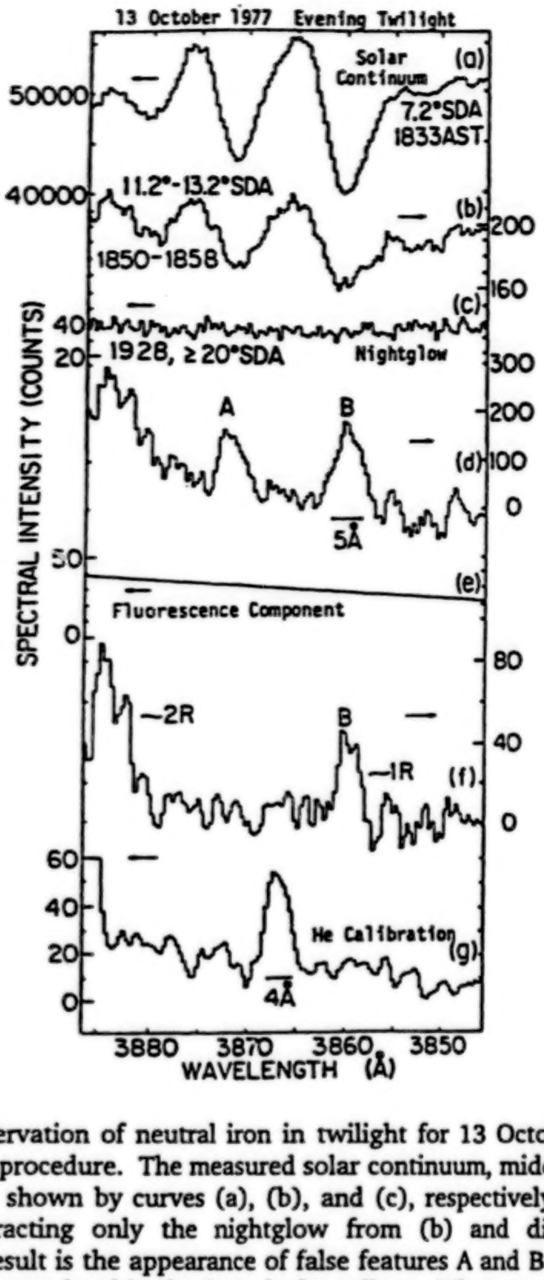


Figure 5 - An observation of neutral iron in twilight for 13 October 1977 illustrating the data reduction procedure. The measured solar continuum, middle twilight spectrum, and nightglow are shown by curves (a), (b), and (c), respectively. Spectrum (d) was obtained by subtracting only the nightglow from (b) and dividing by the solar continuum. The result is the appearance of false features A and B which are inversions of a residual continuum level in the Fraunhofer valleys. When an additional component with the trend given by (e) is also subtracted from (b) before it is ratioed with (a), then the result is shown in spectrum (f). The new feature B is attributed to the Fe resonance line at 3860Å having a width matching a He calibration line shown in the bottom curve. An additional feature in (f) at 3884Å is a portion of the R-branch of the $N_2^+(1\text{-}1)$ 1^{st} Negative band [see Tepley et al., 1981a].

$$\frac{M(\lambda) - C(\lambda)}{M_0(\lambda)} = \frac{aS_0(\lambda) + bF_0(\lambda) + L(\lambda)\delta(\lambda-\lambda_0) - C(\lambda)}{S_0(\lambda) + F_0(\lambda)} \tag{5.3}$$

Since the ratio F_0/S_0 is generally less than 5%, we can expand the function $S_0^{-1}[1+F_0/S_0]^{-1}$ in a binomial series and rewrite equation (5.3) as

$$\begin{aligned}\frac{M(\lambda) - C(\lambda)}{M_o(\lambda)} &= a + \frac{1}{S_0(\lambda)}\,[(b - a)F_0(\lambda) - C(\lambda)] \\ &\quad + \frac{L(\lambda_0)}{S_0(\lambda_0)}\left(1 - \frac{F_0(\lambda_0)}{S_0(\lambda_0)}\right)\end{aligned} \tag{5.4}$$

where we have dropped terms of order S_0^{-2} which are negligible. If we chose C to be identical with $(b-a)F_0$, then equation (5.4) becomes linearly proportional to the line emission amplitude modified by the factor $(1 - F_0/S_0)$, which is close to unity. The wavelength variation of S_0^{-1} is small compared with the instrumental resolution $\Delta\lambda$ so that the line amplitude L can be extracted. To accomplish this, we remove the baseline "a" by subtraction and then multiply by $S_0(\lambda_0)$ to find the line amplitude.

The procedure, then, is to begin the processing with an assumed value for C, and then examine the resultant trial ratios for indications of Fraunhofer structure, which may be inverted if the trail value of C is too small (see Figure 5, curve d). After a number of iterations with different values for C, the processing is complete when the background signal is free of Fraunhofer structure and the width of any spectral line matches the width of the instrumental function as calibrated. These trial and error steps can be quite tedious. If large amounts of spectral data need to be processed routinely, it is best to automate the technique as demonstrated by Swift et al. [1988].

A couple of points about the behavior of the fluorescence component as a function of increasing solar depression angle are worth noting. If the scattered sunlight and the fluorescence decay together, then $b = a$ and $C = 0$, thus, S is proportional to F and it correctly represent the spectral shape of the fluorescence. The value of C required to produce the final result for solar depression angle α can increase as a function of α. If $C > 0$, then $b > a$ and this behavior reflects the fact that the fluorescence decays slower than the scattered light continuum which decreases exponentially at a rate determined by the neutral atmospheric scale height H. This represents a wide and uniform distribution of aerosols in the stratosphere and mesosphere having a scale height greater than H [Giovane et al., 1976]. On the other hand, the situation $C < 0$ is entirely possible. For this last case, $a > b$ and a fluorescence continuum which decreases faster than the scattered light demonstrates a sharper layering of aerosols with a scale height less than H, such as an accumulation of dust at the tropopause [Cadle and Grams, 1975].

The factors a and C are known from the processing. So $b = a + C/F_0$ gives the relative variation of the fluorescence. However, since the magnitude of F_0 is not known, b cannot be determined as a function of time. Furthermore, the $(1 - F_0/S_0)$ correction to the line emission shown in equation (5.4) cannot be applied exactly since F_0 is unknown. Otherwise the emission is correctly determined. Finally, we note that this procedure is correct only if F_0/S_0 is much less than unity.

5.2 Examination of the behavior of fluorescence

For any given spectrum at time $t=t_i$, let $C(t_i)=C_i$, $b(t_i)=b_i$, and $a(t_i)=a_i$, where t_i represents the time at some solar depression angle α_i. At time t_o (that is, for α_o) we have the solar continuum measurement (see equation 5.1). Also at t_o we assume $a_o=b_o=1$, therefore, $C_o=0$.

We then have the relation at $t = t_i$,

$$C_i = (b_i - a_i)\, F_0 \tag{5.5}$$

described earlier, where a_i and C_i are known but b_i and F_0 are unknown. To attempt to determine these parameters exactly we need additional information. We know the shape of the curves describing C_i and a_i from the data processing (cf. Figure 6 in the next section) so the slope of these curves can be found. Differentiating C_i with respect to t_i yields

$$\frac{\partial C_i}{\partial t_i} = F_0\left[\frac{\partial b_i}{\partial t_i} - \frac{\partial a_i}{\partial t_i}\right] \tag{5.6}$$

Partial derivatives are used since there is also a wavelength (λ) dependence. But if we are only concerned with a "short" spectral range in the vicinity of the emission line, then $\partial C_i/\partial\lambda = \partial b_i/\partial\lambda = \partial a_i/\partial\lambda = 0$ and we can replace the partials with total derivatives.

Next, we eliminate F_0 from equation (5.5) and (5.6) to obtain the differential equation for b_i,

$$\frac{db_i}{dt_i} - \left[\frac{1}{C_i}\,\frac{dC_i}{dt_i}\right] b_i + \left[\frac{a_i}{C_i}\,\frac{dC_i}{dt_i} - \frac{da_i}{dt_i}\right] = 0 \tag{5.7}$$

All terms in the brackets are known and can be considered constant at a specific time. However, to obtain a more general solution we must allow C and a to vary with time.

Holding the bracketed terms of equation (5.7) constant at $t=t_i$ we can rearrange the equation to attempt a solution,

$$\frac{d}{dt_i}(b_i - a_i) - \left[\frac{1}{C_i}\,\frac{dC_i}{dt_i}\right](b_i - a_i) = 0 \tag{5.8}$$

Letting $\xi_i = b_i - a_i$, we can separate the variables of equation (5.8) to produce

$$\frac{1}{\xi_i}\,\frac{d\xi_i}{dt_i} = \frac{1}{C_i}\,\frac{dC_i}{dt_i} = K_1 \tag{5.9}$$

where K_1 is the separation constant. This equation simply states that the slope of b-a divided by its magnitude is equal to the rate of change of C divided by its own magnitude at time $t = t_i$, within an arbitrary constant K_1. A quasi-solution to (5.9) might look like

$$\xi_i = (b_i - a_i) = C_i e^{K_1} = C_i K_2 \tag{5.10}$$

where a new constant absorbs K_1 plus the integration constants produced from equation (5.9). Comparing equation (5.10) and (5.5) simply shows that $F_0 = K_2^{-1}$ is a constant which we have assumed in the analysis.

One may think that it might be an easy task to determine this "constant of fluorescence" from the time variations of a_i and C_i by solving simultaneous equations generated at each time t_i. However, for times $t=t_1, t_2, \ldots$, we have the unknown F_0 plus new unknowns b_i, that is, $b_1, b_2, \ldots$ As a result we always have more unknowns than equations that can be formed. Even at $t=t_o$, $a_o=b_o=1$, but $C_o=0$ so K_2 (or F_0) cannot be determined by this method.

The only alternative is to obtain an accurate measurement of F_0 before each twilight experiment. This can be accomplished by viewing the depth of a solar Fraunhofer absorption feature directly by pointing the instrument at the sun, exercising appropriate care, of course. (Direct sunlight is necessary since the use of a diffuser or a white cardboard to reflect light into the instrument will produce an unwanted additional fluorescence of the diffusing material.) After a measurement of the sun is accomplished, point the instrument at the atmosphere near but away from the sun and remeasure the depth of the absorption valley. The percentage difference of the two valleys yields the value of the fluorescence parameter F_0. Using this F_0 and the subsequent variations of a_i and C_i in the twilight analysis, will allow for the determination of b_i, that is, the distribution of the fluorescence with time and corresponding altitude.

5.3 Error accumulation for the difference and ratio methods

We have mentioned that a ratio algorithm is superior to differencing in reducing the errors involved. For differencing, the choice of a small wavelength region (or a single point) for spectral normalization of the solar structure can be a source of large error. This error can be minimized by selecting additional normalization channels and finding the average of the supplementary reduced spectra. We have essentially done this with the ratio technique by taking into account every point in the spectrum when we do the division, that is, the entire spectrum is treated as a single unit.

To illustrate the errors involved in the two methods consider twilight spectral measurements M(t) having error $\Delta M(t)$. We will use t_o to represent the time of measurement of an early twilight spectrum composed mostly of the solar continuum (such as equation 5.1) and t_i to stand for other measurement times during intermediate twilight periods (say, > 5° or 6° solar depression). The difference algorithm uses a scale factor ϵ ($<<1$) applied to the solar structure to remove it by subtraction or

$$M(t_1) \pm \Delta\, M(t_1) - \epsilon[M(t_0) \pm \Delta\, M(t_0)] \tag{5.11}$$

The upper bound of the error is found by making the difference in equation (5.11) as large as possible, or

$$M(t_1) - \epsilon M(t_0) + [\Delta M(t_1) + \epsilon \Delta M(t_0)] \tag{5.12}$$

The expression in brackets is the absolute error of the difference. The relative error in this case is the ratio of the absolute error to the magnitude of the difference. If the errors are small compared with the estimate of the measurement (i.e., $\Delta M << M$) then the relative error is given as

$$\frac{\Delta M(t_1) + \epsilon \Delta M(t_0)}{|M(t_1) - \epsilon M(t_0)|} =$$

$$\frac{M(t_1)}{|M(t_1) - \epsilon M(t_0)|}\left(\frac{\Delta M(t_1)}{M(t_1)}\right) + \frac{\epsilon M(t_0)}{|M(t_1) - \epsilon M(t_0)|}\left(\frac{\Delta M(t_0)}{M(t_0)}\right) \tag{5.13}$$

The terms in the large parenthesis are the relative errors of each component. Each relative error is divided by the weighing factor $|M(t_1) - \epsilon M(t_o)|$ so that the total relative error is not simply the sum of the individual errors. The error can become large since ϵ is ideally chosen to allow $\epsilon M(t_o) \approx M(t_1)$ during data reduction.

In the ratio technique, the upper bound for the error which is similar to equation (5.12) is

$$\frac{M(t_1) + \Delta M(t_1)}{M(t_0) - \Delta M(t_0)} \tag{5.14}$$

We can rationalize this expression, neglect terms of order $(\Delta M)^2$, and rearrange to obtain

$$\frac{M(t_1)}{M(t_0)} + \underline{\frac{M(t_1)}{M(t_0)}\left[\frac{\Delta M(t_0)}{M(t_0)} + \frac{\Delta M(t_1)}{M(t_1)}\right]} \tag{5.15}$$

where we have underlined the absolute error of the quotient. The relative error is the underlined term divided by the magnitude of the quotient which, for the case [illegible]all errors, is simply the expression in brackets.

Now to simplify the expressions, we let $M(t_i)$ be represented by M_i $(i=0,1)$ and $\Delta M_i \equiv \Delta M(t_i) = \sigma$ where the errors σ are assumed to be equal for all spectra. The validity of this assumption is not critical for what we develop below. The ideal value for ϵ in the difference algorithm is M_1/M_o, so the absolute error for differencing (equation 5.12) becomes

$$D = \left[\frac{M_0 + M_1}{M_0}\right]\sigma \tag{5.16}$$

But, for the ratio method the absolute error is written as

$$R = \left[\frac{M_0 + M_1}{M_0^2}\right]\sigma \qquad (5.17)$$

so that the ratio of errors of the two methods is related to the magnitude of the solar continuum, that is, $D/R=M_0$. We note that this is true for all values of ε where $0 < \varepsilon < 1$ because $|M_1|<|M_0|$.

Finally, even though the relative error may not be an adequate representation of the true errors involved, the presence of the pole in the equation generated by the choice of the factor ε at some spectral elements indicates that the absolute errors can become large. On the other hand, there are no poles in the relative error of the ratio method which indicates that the absolute errors are stable. In light of these facts we see that the ratio technique is far superior to the method of differencing in minimizing the errors characteristic of the data analysis.

6. DISCUSSION

In the previous section we suggested that a measurement of a time sequence of optical twilight spectra might contain information about the distribution of aerosols. The amount of correction that is applied to remove the contaminating effects of aerosol fluorescence has a characteristic temporal and wavelength distribution different from the Rayleigh scattered solar continuum. By examining each spectral contribution from middle twilight where the fluorescence component is significant, and the rate at which the magnitudes of these components decay as the solar depression angle increases, a relative scale height might be obtained for the responsible contaminating mechanism or species.

A parameter that can represent the amount of atmospheric fluorescence continuum is the relative amplitude of the required correction with respect to the magnitude of the midtwilight spectra at the wavelength position of a particular resonance line. This signal level can be between a few percent and about 20% of the amplitude of midtwilight spectra. The time progression of the amplitude for the correction generally decreases at a rate less than the solar continuum. For example, in Figure 6 we show a few representative cases of the fluorescence correction amplitude plotted as a function of solar depression angle. These results were derived from several attempts to measure various metallic species of the lower ionosphere during a number of twilight periods at Arecibo. Shown are amplitude plots for different wavelengths. We can obtain some information on the degree of fluorescence even though an emission line is not present in the spectrum. For example, no manganese emissions were ever detected at 4031Å. But the amount of correction required to "flatten" the midtwilight spectra, that is, by removing the solar Fraunhofer structure, can indicate the degree of atmospheric fluorescence near 4000Å.

We suggested that the fluorescence component arises from the presence of aerosols in the lower atmosphere. Basically aerosols can be described as mineral. The reader may have seen museum displays of minerals or meteorites fluoresce under ultraviolet light. The minerals absorb ultraviolet radiation and fluoresce at longer wavelengths.

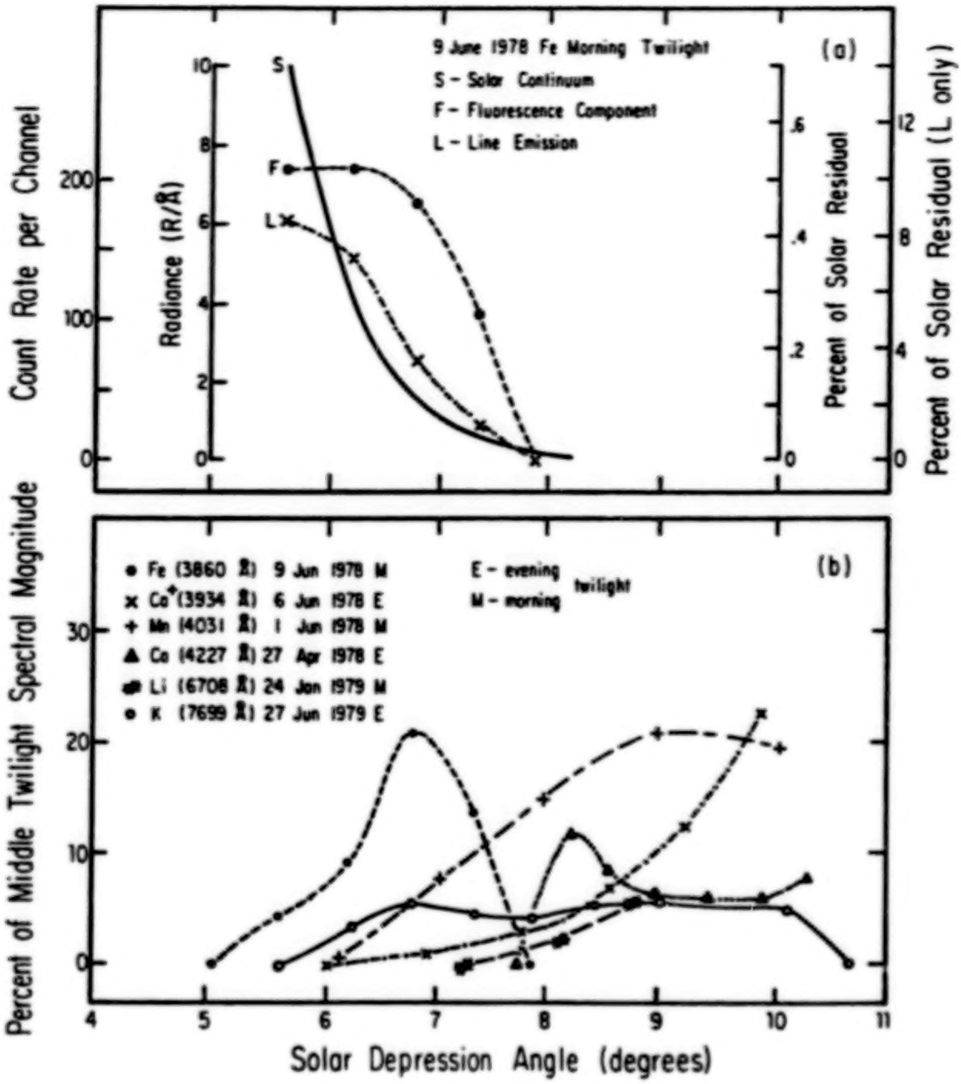


Figure 6 - The temporal variation of the magnitude of the fluorescence component for several wavelengths derived from optical twilight observations at Arecibo. The top panel illustrates the amplitude relationship between the three twilight components, S, F, and L, with respect to the residual level of the solar Fraunhofer absorption (c.f. equation 5.2 in the text) for an Fe observation during a morning twilight [see Tepley et al., 1981a]. Panel (b) shows the relative percentage of the fluorescence component with respect to the magnitudes of the middle twilight spectra which are required to produce flat and featureless spectra, that is, without inverted or non-inverted Fraunhofer absorptions. These results, as a function of solar depression angle, indicate a greater degree of fluorescence in the blue than in the red portions of the spectrum.

Fluorescence is observed in the near UV, blue, and green portions of the spectrum while very few emissions are observed in red. We see from Figure 6 that most results are near the blue portion of the spectrum. But, by comparing the Ca^+ and Fe results to the Li and K fluorescence profiles, there appears to be a greater degree of fluorescence in the blue than in the red, at least qualitatively. We should mention that these data do not adequately cover the expected seasonal effects inherent in the implied aerosol fluorescence. We will discuss these aspects further.

In twilight data analysis we have stressed the need to account for the continuum resulting from the fluorescence of aerosols which are bathed in direct solar or scattered radiation. These aerosols are in the form of dust, NO_2, or pollutants found in the stratosphere and troposphere. The concentrations of aerosols may be low but the amount of scattering in the lower atmosphere is high [Kompaniets, 1965]. The continuum emitted by aerosols when illuminated by the sun will always contain a small fraction of the scattered solar continuum. If the relative fraction of the background continuum due to fluorescence were to remain constant throughout twilight, then the Fraunhofer structure can be removed by either scaling and subtraction [Broadfoot, 1967] or rationing [Noxon et al., 1979]. Furthermore, as the example of Figure 6 illustrated, if the decrease in the amplitude of the fluorescence continuum as a function of solar depression angle is less than the decrease in the scattered solar continuum, then this behavior indicates that the scale height for the distribution of aerosols is greater than the atmospheric scale height. This is expected if the concentrations of aerosols in the stratosphere (and mesosphere) are in the form of broad layers or wide distributions [Rosen, 1969; Link, 1973]. The differencing technique used to remove Fraunhofer structure assumes that the relative intensity distribution of the background continuum will not vary throughout twilight. When the fluorescence decays at a different rate than the scattered light continuum, and it is not taken into account, the effect of normalization by scaling can produce an enhanced emission rate of an airglow line.

We estimate that what we call the fluorescence component of twilight is very similar (if not identical) to the Ring effect found in dayglow observations [Grainger and Ring, 1962; Hunten, 1970; Chanin, 1975]. These studies attributed the Ring effect to albedo changes during solar transit or to rotational Raman scattering of molecules such as N_2 and O_2 [Brinkmann, 1968; Noxon et al., 1979]. The percentage of this scattering has been measured to usually increase or remain constant with increasing solar zenith angle. Barmore [1975] however, reported both increasing and decreasing scattering as the solar zenith angle increased using high resolution dayglow measurements of the 6300Å Fraunhofer absorption line. He was able to eliminate other possible mechanisms for the scattering as being of little significance, and concluded that the variations observed were due to random variations in the amount of fluorescing material. For example, the surface of the Earth does indeed fluoresce and this may be important when the atmospheric turbidity is low. However, the presence of the Ring effect after sunset indicates that there is a fluorescence in the stratosphere. Kattawar et al. [1981] on the other hand, favor Brinkmann's proposed mechanism for the filling-in of Fraunhofer lines and extended the theory to include inelastic Rayleigh-Brillouin scattering together with rotational Raman scattering.

From observations at low latitudes we have found that the magnitude of the fluorescence component of twilight spectra is variable from day to day displaying both increasing and decreasing trends with increasing solar depression angle. Its relative magnitude was observed to vary by as much as 20% within a single twilight. There is also a great variability in the dust and aerosol concentrations with geographic location, in particular for the lower latitudes [Prospero et al., 1970; Carlson and Prospero, 1972]. Because of this daily variability of aerosols, it is also unwise to use the solar continuum from one twilight to reduce the spectra of another twilight, even though care is maintained to set up identical experimental conditions. However, by applying proper data analysis procedures as outlined in this report, the measurement of weak emissions during twilight can contribute a great deal to the understanding of atmospheric composition and the relative distribution of aerosols in the middle and lower atmosphere.

ACKNOWLEDGEMENTS

I am very grateful to Ms. Carmen Torres for her proficiency and patience during the preparation of this report. I would also like to thank Dr. Doug Torr for allowing me to present some of the work of his group (figure 1) before publication. Results shown in some of the other figures were from observations at the Arecibo Observatory which is operated by Cornell University under a cooperative agreement with the National Science Foundation.

REFERENCES

Barbier, D., Recherches sur la raie 6300 de la luminescence atmosphérique nocturne, *Ann. Geophys., 15,* 179-217, 1959.

Barmore, F. E., The filling-in of Fraunhofer lines in the day sky, *J. Atmos. Sci., 32,* 1489-1493, 1975.

Brinkmann, R. T., Rotational Raman scattering in planetary atmospheres, *Astrophys. J., 154,* 1087-1093, 1968.

Broadfoot, A. L., Twilight Ca^+ emission from meteor trails up to 280 km, *Planet. Space Sci., 15,* 503-514, 1967.

Broadfoot, A. L., and A. E. Johanson, Fe(3860Å) emission in the twilight, *J. Geophys. Res., 81,* 1331-1334, 1976.

Broadfoot, A. L., and D.M. Hunten, N_2^+ emission in the twilight, *Planet. Space Sci., 14,* 1303-1319, 1966.

Brown, T. L., The chemistry of metallic elements in the ionosphere and mesosphere, *Chem. Rev., 73,* 645-667, 1973.

Cadle, R. D., and G. W. Grams, Stratospheric aerosol particles and their optical properties, *Rev. Geophys. Space Phys., 13,* 475-501, 1975.

Carlson, T. N., and J. M. Prospero, The large-scale movement of Saharan air outbreaks over the northern equatorial Atlantic, *J. Appl. Meteor., 11,* 283-297, 1972.

Chamberlain, J. W., Resonance scattering by atmospheric sodium-I Theory of the intensity plateau in the twilight airglow, *J. Atmos. Terr. Phys., 9,* 73-89, 1956.

Chamberlain, J. W., *Physics of the Aurora and Airglow,* Academic Press, New York, NY, 1961.

Chamberlain, J. W., *Theory of Planetary Atmospheres: An Introduction to Their Physics and Chemistry,* 1st edition, Academic Press, New York, NY, 1978.

Chanin, M. L., Filling in of the Fraunhofer lines by scattering on the ground, *J. Geophys. Res., 80,* 2859-2862, 1975.

Clemesha, B. R., Lidar studies of the alkali metals, MAP Handbook, *13,* 99-112, 1984.

Donahue, T. M., R. Resnick, and V. R. Stull, Distribution in the upper atmosphere of sodium atoms excited by sunlight, *Phys. Rev., 104,* 873-879, 1956.

Dufay, M., Étude de l'emission de la molécule d'azote ionisée et de l'atome neutre d'azote au crépuscule, *Ann. Phys., 8,* 813-862, 1959.

Evans, W. F. J., and E. J. Llewellyn, Measurements of mesospheric ozone from observations of the 1.27μ band, *Radio Sci., 7,* 45-50, 1972.

Evans, W. F. J., H. C. Wood, and E. J. Llewellyn, Ground-based photometric observations of the 1.27μ band of O_2 in the twilight airglow, *Planet. Space Sci., 18,* 1065-1073, 1970.

Fedorova, N. I., On emission λ 10830Å in aurorae, *Planet. Space Sci., 5,* 75, 1961.

Fricke, K. H., and U. von Zahn, Mesopause temperatures derived from probing the hyperfine structure of the D_2 resonance line of sodium by lidar, *J. Atmos. Terr. Phys., 47,* 499-512, 1985.

Gadsden, M., Antarctic twilight observations, I. Search for metallic emission lines, *Ann. Geophys., 25,* 667-677, 1969.

Gardner, C. S., and J. D. Shelton, Density response of neutral atmospheric layers to gravity wave perturbations, *J. Geophys. Res., 90,* 1745-1754, 1985.

Giovane, F., D. W. Schuerman, and J. M. Greenberg, The solar occultation technique for remote sensing of particulates in the Earth's atmosphere, 2. Skylab results of a 48-km aerosol layer, *J. Geophys. Res., 81,* 5383-5388, 1976.

Grainger, J. F., and J. Ring, Anomalous Fraunhofer line profiles, *Nature, 193,* 762, 1962.

Granier, C., J. P. Jegou, and G. Megie, Iron atoms and metallic species in the Earth's upper atmosphere, *Geophys. Res. Lett., 16,* 243-246, 1989a.

Granier, C., J. P. Jegou, and G. Megie, Atomic and ionic calcium in the Earth's upper atmosphere, *J. Geophys. Res.,* in press, 1989b.

Granier, C., and G. Megie, Daytime lidar measurements of the mesospheric sodium layer, *Planet. Space Sci., 30,* 169-177, 1982.

Henriksen, K., G. G. Sivjee, and C. S. Deehr, Winter enhancement of atomic lithium and sodium in the polar upper atmosphere, *J. Geophys. Res., 85,* 5153-5156, 1980.

Hunten, D. M., Transmission functions for twilight studies, *J. Atmos. Terr. Phys., 24,*333-338, 1962.

Hunten, D. M., Spectroscopic studies of the twilight airglow, *Space Sci. Revs., 6,* 493-573, 1967.

Hunten, D. M., Surface albedo and the filling-in of Fraunhofer lines in the day sky, *Astrophys. J., 159,* 1107-1110, 1970.

Kattawar, G. W., A. T. Young, and T. J. Humphreys, Inelastic scattering in planetary atmospheres, 1. The ring effect with aerosols, *Astrophys. J., 243,* 1049-1057, 1981.

Kirchhoff, V. W. J. H., and B. R. Clemesha, Atmospheric sodium measurements at 23°S, *J. Atmos. Terr. Phys., 35,* 1493-1498, 1973.

Kompaniets, A. I., Aerosol light scattering indicatrices in the free atmosphere at heights of the order of 10 km, *Izv. Atmos. Oceanic Phys., 1,* 335-338, 1965.

Kwon, K. H., C. S. Gardner, D. C. Senft, F. L. Roesler, and J. Harlander, Daytime lidar measurements of the tidal winds in the mesospheric sodium layer at Urbana, Illinois, *J. Geophys. Res., 92,* 8781-8786, 1987.

Link, F., Aerosols and particles, in *Physics and Chemistry of Upper Atmospheres*, B. M. McCormac, ed. D. Reidel, Dordrecht, Holland, pp. 34-40, 1973.

Lowe, R. P., K. L. Gilbert, and R. J. Niciejewski, Twilight intensity variation of the infrared hydroxyl airglow, MAP Handbook, *10,* 176-178, 1984.

Lytle, E. A., and D. M. Hunten, Observations of N_2^+ twilight and sunlit aurora, *Can. J. Phys., 40,* 1370-1384, 1962.

McElroy, M. C., and D. M. Hunten, A method of estimating the Earth albedo for dayglow measurements, *J. Geophys. Res., 71,* 3635-3638, 1966.

Megie, G., F. Bos, J. E. Blamont, and M. L. Chanin, Simultaneous nighttime lidar measurements of atmospheric sodium and potassium, *Planet. Space Sci., 26,* 27-35, 1978.

Meriwether, J. W., Jr., D.G. Torr, J. C. G. Walker, and A. O. Nier, The $O^+(^2P)$ emission at 7320Å in twilight, *J. Geophys. Res., 83,* 3311-3319, 1978.

Meriwether, J. W., Jr., Measurement of weak airglow emissions with a scanning spectrometer, *Planet. Space Sci., 27,* 1221-1232, 1979.

Noxon, J.F., Twilight enhancement in $O_2(b^1\Sigma_g)$ airglow emission, *J. Geophys. Res., 80,* 1370-1373, 1975.

Noxon, J. F., E. D. Whipple, and R. S. Hyde, Stratospheric NO_2, 1. Observational methods and behavior at mid-latitude, *J. Geophys. Res., 84,* 5047-5065, 1979.

Noxon, J. F., A study of the 6300Å oxygen line in the day airglow, *J. Geophys. Res., 69,* 3245-3255, 1964.

Pick, D. R., E. J. Llewellyn, and A. Vallance Jones, Twilight airglow measurements of the OH and O_2 bands by means of balloon-borne instruments, *Can. J. Phys., 49,* 897-905, 1971.

Prospero, J. M., E. Bonatti, C. Schubert, and T. N. Carlson, Dust in the Caribbean atmosphere traced to an African dust storm, *Earth Planetary Sci. Lett., 9,* 287-293, 1970.

Rosen, J. M., Stratospheric dust and its relationship to the meteoric influx, *Space Sci. Res., 9,* 58-59, 1969.

Swider, W., Ionic and neutral concentrations of Mg and Fe near 92 km, *Planet. Space Sci., 32,* 307-312, 1984.

Swift, W., D. G. Torr, M. R. Torr, C. Hamilton, and P. Bhatt, Spectral extraction from twilight airglow, paper at 1988 CEDAR Summer Workshop, Boulder, CO, 1988.

Tepley, C. A., J. W. Meriwether, Jr., J. C. G. Walker, and J. D. Mathews, Observations of neutral iron emission in twilight spectra, *J. Geophys. Res., 86,* 4831-4835, 1981a.

Tepley, C. A., J. D. Mathews, J. W. Meriwether, Jr., and J. C. G. Walker, Observations of the Ca^+ twilight airglow from intermediate layers of ionization, *J. Geophys. Res., 86,* 7781-7786, 1981b.

Torr, D. G., M. R. Torr, J. W. Meriwether, Jr., and R. G. Burnside, Measurements of the metastable $N^+(^1S)$ 5755Å emission in the twilight thermosphere, *J. Geophys. Res., 86,* 2314-2316, 1981.

Torr, D. G., W. A. Abdou, M. R. Torr, R. G. Burnside, and J. W. Meriwether, Jr., Measurements of the metastable $N^+(^1D)$ 6584Å emission in the twilight thermosphere, *J. Geophys. Res., 88,* 3190-3196, 1983.

Torr, M. R., D. G. Torr, P. Bhatt, W. Swift, and H. Dougani, Ca^+ emission in the sunlit ionosphere, *J. Geophys. Res.,* in press, 1989.

Vallance Jones, A., and A. W. Harrison, Calcium and oxygen in the twilight airglow, *Ann. Geophys., 14,* 179-185, 1958.

von Zahn, U., and R. Neuber, Thermal structure of the high latitude mesopause region in winter, *Beitr. Phys. Atmosph., 60,* 294-304, 1987.

Chapter 9

OPTICAL AND RADAR TECHNIQUES APPLIED TO CHEMICAL RELEASE "ACTIVE EXPERIMENTS" IN THE IONOSPHERE/THERMOSPHERE SYSTEM

Michael Mendillo, Jeffrey Baumgardner and Peter J. Sultan
Center for Space Physics
Boston University
Boston, MA 02215 USA

I. INTRODUCTION TO ACTIVE EXPERIMENTS

I.1 Background.

The term "Active Experiments" refers to a relatively new class of space physics experimentation in which a known perturbation is applied to a natural system in order to study its characteristics through observations and analysis of the induced effects. Such techniques have been applied to the Earth's thermosphere/ionosphere, the magnetosphere and solar wind. There are four general types of stresses used as causative agents in active experiments: (1) neutral gas releases, (2) cold plasma injections, (3) ion and electron beams, and (4) electromagnetic wave injections. Areas (1), (2), and (3) can be handled via sounding rocket or satellite-borne payloads, while (4) can utilize either powerful groundbased transmitters or spaceborne injection methods.

Given the relatively modest scientific payloads that can be carried into earth orbit, the spatial extent and temporal duration associated with active experimentation are small and brief in comparison to those associated with natural, solar-terrestrial-type disturbances. For example, chemical release experiments in the ionosphere produce disturbances spanning 10's to 100's of kilometers that last a few minutes to perhaps a few hours. Ionospheric storms, on the other hand, are global in scale and last for days. Similarly, barium plasma injections in the ionosphere or magnetosphere result in observable effects that typically last for periods less than the duration of a substorm (~ few hours). Ionospheric heating effects generally last as long as the heating is applied, often in the 10's of minutes domain. The end result is that active experiments require a comprehensive set of multi-diagnostic instruments that can handle the small time constants associated with the induced effects. These pose no particular problems for in-situ probes that sample the regions close to the perturbation; sounding rocket and satellite diagnostics routinely provide high data sets. However, when groundbased radars and optical systems are used in active experiments, the normal modes used for sampling geophysical effects must often be modified to cope with the reduced spatial and temporal time scales of active experiments.

In the following sections, we concentrate on the chemical release experiments used in ionospheric investigations, namely, neutral gas and cold plasma injection experiments. For a general overview of all types of active experiments, the reader is referred to the proceedings of several symposia on Active Experiments in Space held in recent year [Burke, 1983; Haerendel and Mendillo, 1988; Mendillo, 1989; and Brenning, 1989].

I.2. Types of Chemical Release Experiments.

Chemically-induced ionospheric modification experiments involve either the enhancement of ambient plasma densities or their depletion. In the former case, the most widely used technique involves the release of barium vapor into the ionosphere [Foppl et al., 1967; Haerendel, 1976]. Barium is an element that is ionized by solar ultraviolet with a time constant of ~ 28 seconds. The Ba^+ ion has bright emission lines at 4554 and 4934Å, with less intense lines at 5854, 6142, and 6497Å. In its neutral state, Ba has a bright emission line at 5535Å. Barium release experiments are conducted under post-sunset or pre-sunrise conditions using geometries required for darkness on the ground (to permit optical observations) but UV illumination in the ionosphere (to permit optical excitation). Under such conditions, observations can be made for 10's of minutes, depending upon the release altitude and ground site locations.

A Ba^+/e^- plasma responds to electrodynamical effects and thus optical observations can be used to trace ambient electric fields associated with ionospheric-magnetospheric coupling, space plasma instabilities, critical ionization velocity (CIV) effects, and many other phenomena. Again, the reader is referred to the special issues quoted above and to the references contained therein.

Strontium (Sr) vapor has been used in conjunction with Ba release experiments (contained as an impurity or "dopant"), or as an independent species released into the ionosphere/thermosphere. Sr has the property of not being readily ionized ($\tau_i \sim 1$ hour), but is excited in its neutral state by sunlight, and thus it can be used to illuminate the evolution of a gas not subject to electrodynamical forces. It can, for example, trace neutral wind patterns or map diffusion processes. The Sr emission line typically observed is at 4607Å; when Sr^+ is formed, observations can be made at 4078Å.

Many Ba^+ and Sr release experiments are not true active experiments in that they do not modify ambient systems, but rather illuminate or trace existing features. The same can be said for releases of other species such as trimethyl aluminum (TMA), cesium, calcium, etc. These experiments do, however, have the common feature with active experiments of offering a short duration, visible light signal that is observed at a wavelength not confused with regular atmospheric emissions. With often very bright lines to observe, a wide variety of techniques can and have been used to record such emissions, ranging from standard film cameras to specialized video and CCD-based digital imaging systems [Heppner et al, 1981; Stenbaek-Nielsen et al., 1984; Wescott et al., 1986].

The second type of chemical release experiment involves the release of molecules that react very rapidly with ambient ionospheric plasma (Mendillo and Forbes, 1982; Bernhardt, 1987a). These highly reactive rapors (e.g., H_2, CO_2, H_2O, SF_6, CF_3Br, $Ni(CO)_4$) modify the ionosphere via recombination chemistry (plasma neutralization) that leaves the ionospheric F-region plasma in a depleted ("ionospheric hole") state. As part of the photo-chemical chains in the modification process, there is emission of light from the oxygen atoms produced by the neutralization of the ambient O^+/e^- plasma. These enhanced airglow lines (typically at 6300, 7774 and 8446Å) can be used to monitor the chemical reactions and neutral dynamics of the system. Recent reviews of the mechanisms, goals and status of neutral gas induced active experiments have been given by Mendillo (1988) and Bernhardt (1987a).

The diagnostic challenges presented by neutral gas induced effects center on (1) the reduced signal-to-noise levels for a radar probing the remnant plasma in an ionospheric hole and (2) the short duration, sub-visual airglow enhancements at ambient emission wavelengths that accompany the formation of ionospheric holes. These topics are treated separately in the following sections. They have an obvious relevance to plasma enhancement active experiments where stronger signals would require less integration time for radar and imaging systems.

II. Incoherent Scatter Radar Techniques Applied to Active Experiments

II.1. Background

Incoherent scatter radar is one of the more important and useful observational tools for studying the physics of the upper atmosphere [Evans, 1969,1975]. Physically, it is based on the fact that an upwardly propagating radio wave that is higher in frequency than the plasma frequency anywhere along its path will be weakly re-radiated by the free electrons of the ionosphere. This incoherent scatter, or "Thomson" scattering, provides a small but measureable return signal from a probing radar beam. Because of the weakness of the scattered signal, high powered transmitters and large collecting areas are required for adequate detection.

The returned incoherent scatter signal may be analyzed to determine a wide variety of ionospheric parameters. Of principal interest in active experiments are:

N_e - local electron density. N_e is proportional to the area under the curve of the returned spectrum, and is typically observed up to h ~ 1000 km.

V_i - ion sound speed: derived from the overall width of the spectrum, or from the location of the spectral peaks.

T_i - ion temperature: derived from V_i if the ion masses are known.

T_r - ratio of T_e/T_i: derived from the shape of the returned spectrum.

T_e - electron temperature: calculated from T_r, once T_i is determined.

V_d - large-scale drift of the ionospheric plasma: determined from an observed doppler shift of the entire spectrum.

Currents - differential drifts of electrons and ions can be detected from the presence of different heights of the up- and down-shifted spectral wings [Rosenbluth and Rostoker, 1962].

Composition - mixtures of different ions can be inferred from changes in the shape of the returned spectrum [Moorcroft, 1964].

Simultaneous changes in several different parameters can have a complex effect on the spectral signal. However, in the natural ionosphere, changes in individual spectral parameters occur over a limited altitude range, making it possible to deconvolve individual effects by making reasonable assumptions as to which parameters are changing, and which remain constant. Thus the scattering from each altitude region can be described in terms of changes in only two or three important parameters.

A typical incoherent scatter analysis program determines ionospheric parameters by calculating theoretical spectra which most clearly match the observed spectra. The various physical parameters controlling the spectrum shape are adjusted to provide for the best least-squares fit between theory and data. An approximation of one or more parameters, such as the ion composition or T_e/T_i profile, can often be made so the other parameters can be derived without ambiguity. Traditional incoherent scatter radar analyses often begin by assuming an O^+/e^- plasma of $T_e/T_i = 1$ near the F_2 peak at night. Additionally, one or more parameters can be pre-determined from an additional data source (e.g. obtain f_oF2 from an ionosonde, or include data from measurements of the incoherent scatter plasma line).

Incoherent scatter radar is a unique and ideal diagnostic instrument for the measurement of ionospheric parameters during active experiments. If the antenna is quickly steerable, there is an added flexibility for observations under rapidly changing circumstances (e.g., sudden changes in rocket trajectory). Observations over time can yield an excellent spatial/temporal picture of the changing conditions in the ionosphere during chemical releases.

The first use of a steerable incoherent scatter radar to observe an artificial, short-duration event took place in 1979 when the Millstone Hill Incoherent Scatter Facility (42.62°N, 71.49°W) measured an ionospheric disturbance to the east of the Kennedy Space Center following the launch of NASA's HEAO-C satellite by an Atlas/Centaur rocket [Wand and Mendillo, 1984]. A chemically-induced rapid recombination of F-region plasma caused by the rocket's exhaust cloud of H_2 and H_2O molecules was monitored by the 46 m diameter Millstone Hill radar, and the spatial, temporal, and dynamical morphologies of a large-scale 'ionospheric hole' were observed for the first time. The incoherent scatter studies of the rocket launch effects showed that reliable altitude profiles of electron densities, plasma temperatures and bulk plasma drift speeds can be obtained during active experiments.

During the HEAO-C event, the main difficulties with the observations resulted from the large distance between the perturbation and observation sites. As described in Evans (1969), radar range resolution (ΔS) depends upon pulse duration (Δt) via $\Delta S = \frac{1}{2}c\Delta t$ where c is the speed of light. In order to achieve acceptable signal to noise (S/N) levels for the radar returns for the HEAO-C case, long pulse lengths (0.5 ms) had to be employed, thus limiting the range resolution to relatively large values (75 km). Integration time was set at 20 sec for each data record. Such large space and time resolutions are not desirable for observations of effects comparable in size to a single pulse length, or to events changing on minute time scales. These resolutions were adequate for the HEAO-C observations since the Centaur exhaust cloud spanned 100's of km and the event lasted for several hours. However, since a typical chemical release cloud from a sounding rocket payload reaches an approximate maximum radius of ~ 50 km, an improved spatial resolution would be required for such experiments.

An improvement in range resolution to 10-20 km would require 0.1 ms pulses. Unfortunately such short pulses unfavorably impact the S/N levels, and require a longer integration time to obtain usable spectra (i.e. ones from which reliable spectral parameters, N_e, T_e, T_i, V_d, and composition, can be obtained). The advantages of using an incoherent scatter radar as a diagnostic tool quickly deteriorate if the length of integration time needed approaches the time-scales of the event being monitored. In general, attempts to minimize spatial and temporal resolution values conflict with the simultaneous desire to maximize the information obtained from spectral analysis.

A further active experiment observed by the Millstone Hill radar was during the Spacelab-2 mission in 1985 when the Space Shuttle's OMS engines were used to generate an ionospheric hole above New England [Mendillo et al., 1987]. The magnitude of the Spacelab-2 electron density depletion was much smaller than the earlier (HEAO-C) rocket-induced hole because of the smaller shuttle engines and the relatively weak solar minimum ionosphere. A relatively long radar pulse length (0.3 ms) and 20 second integrations were used to maximize the power return to obtain accurate measurements of the electron density, thus determining the trade-off between power and spectral resolution.

During the same Spacelab-2 mission, an engine burn was conducted over the Arecibo Incoherent Scatter Radar Facility in Puerto Rico [Bernhardt et al., 1988]. In this experiment, the 305 m diameter Arecibo dish was able to provide 600 m spatial resolution, using shorter 0.004 ms pulse lengths, and 5.6 second integration data records. Clearly, the Arecibo radar owing to its great size, does not suffer from the same power/spectral resolution trade-off that the smaller Millstone Hill dish does. However, the Arecibo dish has greatly limited steerability, and is generally restricted to observations within 20° of zenith, a potential handicap for use in off-zenith active experiments. For smaller antennas, the lag-profile analysis techniques recently developed for the Millstone Hill radar will help in improving range and time resolution. A 15 km resolution is now possible, representing a significant improvement over the previous 40 km value [Holt et al., 1988]. Thus the trade-off between power return and spectral accuracy faced during the HEAO-C and Spacelab-2 experiments can be relaxed.

II.2. Incoherent Scatter Radar as an Ion Mass Spectrometer.

A problem unique to active experiments is that the composition of the observed ionospheric plasma is changed by the released species, and the radar is often faced with ions not usually present in the natural ionosphere. A spectral analysis program that is not expected to handle unusual ions will misinterpret the presence of those ions as changes in other parameters (i.e., temperature).

The IS spectrum for a mixture of ions is a superposition of the spectra for the individual ions, smoothed out by the Coulomb interactions (Moorcroft 1964). The presence of a mixture will be most clearly revealed in the spectral shape when there is a large difference in the ion masses.

The ions created in the HEAO-C and Spacelab-2 releases [OH^+ (17 amu), H_2O^+ (18 amu), and O_2^+ (32 amu)] were not notably different in mass from naturally occurring ionospheric ions [O_2^+ (32 amu), O^+ (16 amu)]. The effects of such ions on the shape of an IS spectrum are small, and their presence is not easily inferred. However, some active experiments use heavy elements never found in the ionosphere, even under extreme circumstances. Over the years, a variety of experiments have been conducted using Ba (137 amu) releases, and a much smaller number using clouds of SF_6 (146 amu). The planned NICARE-1 and -2 experiments will release CF_3Br and $Ni(CO)_4$, which will produce the ions Br^- (80 amu) and $Ni(CO)_3^-$ (143 amu) [Bernhardt, 1987a].

To illustrate the effect of heavy positive ions on incoherent scatter spectra, a simulation of the spectrum associated with a barium plasma (Ba^+/e^-) in the F-region (O^+/e^-) has been conducted by Sultan et al. [1989]. The main effects are to pull the spectral wings in, and to make the peaks higher (Figure 1). The situation is analagous to the effect of H^+ on the spectrum of an O^+/e^- plasma, where the H^+ pulls the spectral wings outwards and lowers the peaks. In both instances two positive ions of very different masses are being mixed. For a given electron density, as the concentration of one ion is increased, the concentration of the other must decrease in order to keep the plasma electrically neutral. The change in spectral shape as the concentration of heavy positive ions increases is a diminishing of the spectral wings of the lighter ion, while the peaks of the heavier ion become more prominant. The total area under the spectral curve is constant since the background electron density is not changing.

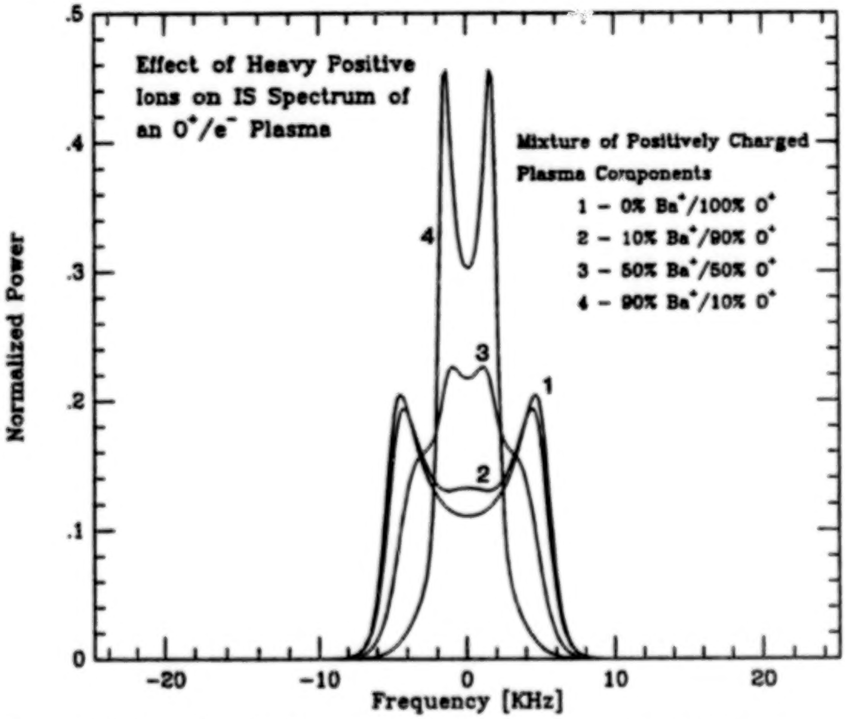


Figure 1. Theoretical incoherent scatter radar power spectra for different concentrations of heavy positive ions (Ba^+, 137 amu) on an O^+/e^- plasma ($T_e/T_i = 2.$). Curves 1 through 4 represent 0%, 10%, 50% and 90% Ba^+, respectively.

Figure 2 illustrates the effect of different concentrations of heavy negative ions (SF_6^-, 146 amu) on the IS spectrum of an O^+/e^- plasma. Such plasmas, containing both heavy negative, and lower-mass positive ions, were created during the NASA/Boston University SPINEX campaign through the release of SF_6 at F-region altitudes. SF_6, which has a high electron affinity, quickly attaches an ambient electron to itself ($SF_6 + e^- \longrightarrow SF_6^-$) creating a heavy, negatively charged ion. Since the newly formed plasma mixture must also be electrically neutral, the number of positively charged components must equal the number of negatively charged components, so that $N(O^+) = N(SF_6^-) + N(e^-)$. For the previous case of heavy positive ions, the charge neutrality relation was $N(O^+) + N(Ba^+) = N(e^-)$.

As was noted in Mendillo and Forbes (1982), the presence of heavy negative ions pulls the spectral wings outwards, and makes the shoulder features sharper and more peaked. An additional central hump is also added. The overall effect is similar to the spectral changes brought on by an increase in the T_e/T_i ratio (Figure 3), with the exception of the new central hump. This new peak, which actually appears to be two low frequency peaks very close to the transmitted radar beam frequency, has an additional mass dependency, and is more pronounced the heavier the ion is. The total area under the spectral curve does decrease in this case, since the total number of electrons available to scatter the radar beam has been reduced due to their conversion to the negatively charged ions.

If such spectral changes are taken into account by the radar analysis routine, an incoherent scatter radar can be used as a de facto mass spectrometer to track the spatial and temporal development of a heavy negative ion plasma cloud created during an active experiment. For a nighttime experiment, the overall strategy is to assume that the ambient temperature and composition of the background ionospheric plasma does not change over the short duration of the experiment, and that the new plasma from the ionized components of the chemical release quickly reaches temperature equlibrium. The temperature structure of the ionosphere is then fixed to a pre-event average, and the spectral analysis program is modified to recognize heavy negative ion effects. The analysis program will then view any spectral changes as being caused by changes in composition, and not due to changes in temperature. A program that does not take the presence of heavy negative ions into account will erroneously attribute the effect of the ions to a temperature change. Thus the ion mass spectrometer capabilities of an incoherent scatter radar can be exploited.

The presence of 10% or more SF_6^- has been successfully detected in the SPINEX-2 (1986) data set [Sultan et al, 1989]. Analysis of the SPINEX-1 (1984) and AFGL IMS (1983) SF_6 release data sets is underway. Incoherent Scatter Mass Spectrometry (ISMS) of heavy negative ions will be further tested during the NASA CRRES SF_6 (1990), and NASA/NRL NICARE releases of CF_3Br and $Ni(CO)_4$ (1989,1990). An incoherent scatter scan has yet to be performed on a heavy positive ion release, but several opportunities will occur during the CRRES experiments over Arecibo in 1991.

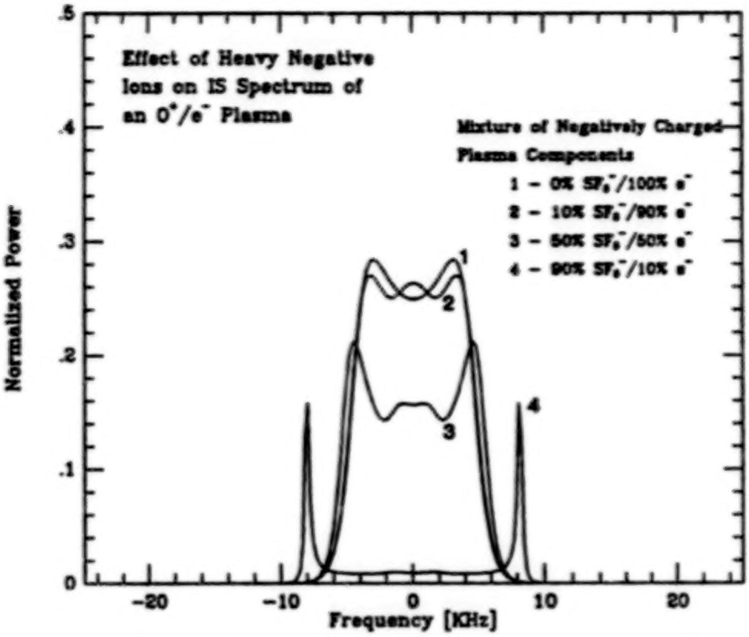


Figure 2. Theoretical incoherent scatter radar power spectra for different concentrations of heavy negative ions (SF_6^-, 146 amu) on an O^+/e^- plasma ($T_e/T_i = 1.$). Curves 1 through 4 represent 0%, 10%, 50% and 90% SF_6^-, respectively.

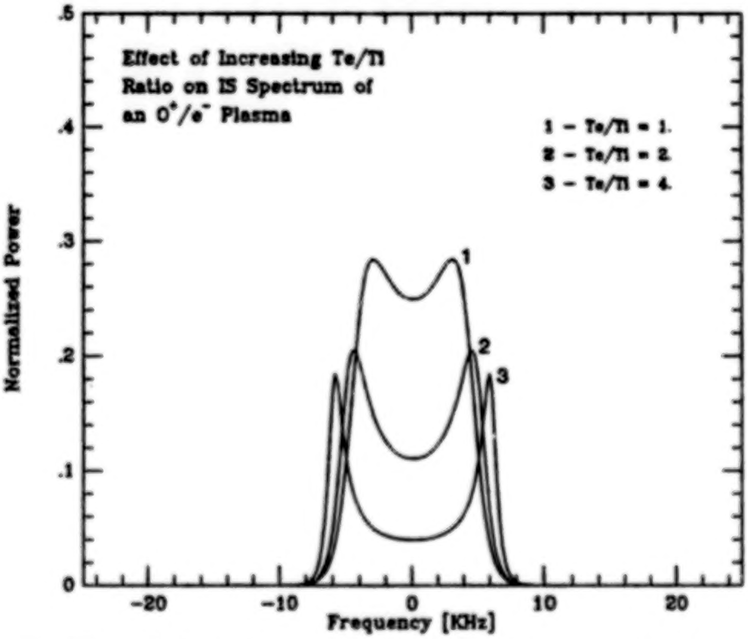


Figure 3. Theoretical incoherent scatter radar power spectra for a pure O^+/e plasma for different values of the T_e/T_i ratio. Curves 1 through 3 represent T_e/T_i = 1., 2, and 4. respectively.

III. Low-Light-Level Optical Instrumentation Applied to Active Experiments

III.1. Background

Optical instrumentation using image intensifiers has been applied to a wide variety of geophysical and space plasma phenomena. A film-based image intensified system was first applied to ionospheric hole airglow studies [Mendillo and Baumgardner, 1982a] using a camera developed for aeronomic studies [Mendillo and Baumgardner, 1982b]. A CCD-based digital camera system was later developed [Baumgardner and Karandanis, 1984] and applied successfully to ionospheric modification studies [Mendillo et al., 1987]. Alternate CCD-based systems were developed and used in several ionospheric and magnetospheric chemical release experiments [Bernhardt et al., 1988, 1987b], and in ionospheric heating experiments [Bernhardt et al., 1989].

Each of the systems described in the references cited successfully addressed the fundamental issues faced by optical diagnostics devoted to active experiments: (1) weak signals at ambient atmospheric emission lines; (2) duty cycle concerns relating length of integration times and image storage times to the time scale of effects to be observed; (3) operational choices concerning dynamic range levels or image intensifier phosphor persistence ("afterglow") effects.

Fabry-Perot systems have been used extensively in aeronomic research, but only recently has the technique been applied to active experiments. A Fabry-Perot system used in high altitude barium release observations was described by Mende and Harris (1982). The AMPTE program of chemical releases included Fabry-Perot measurements of barium released into the geomagnetic tail [Mende et al., 1989], as well as into the solar wind [Rees et al., 1986]. The Doppler Imaging System (DIS) used in the latter case is described in Rees et al., (1984). Opportunities to apply these techniques to releases in the thermosphere-ionosphere system will occur during the NASA CRRES missions scheduled for 1990-91.

In the following sections, brief descriptions will be given of two new instruments that are well suited for observations of active experiments in the ionosphere: a low-light-imager and a meridional imaging spectrograph.

III.2. The CEDAR Imaging System.

As part of a U.S. National Science Foundation initiative in atmospheric sciences, the CEDAR ("Coupling, Energetics and Dynamics of Atmospheric Regions") Program has sought to establish instrumentation and programmatic upgrades in all areas of aeronomic research (CEDAR, 1986). One of the core instruments developed for optical diagnostics was a state-of-the-art imaging system. The CEDAR Class-I imager is expected to advance ground-based capabilities for upper atmospheric research in two fundamental ways: (1) provide high quality, fully-calibrated, all-sky images at multiple wavelengths for research in optical aeronomy, (2) provide a facility class instrument to operate in campaign mode in conjunction with simultaneous observations by incoherent scatter radar, Fabry-Perot, LIDAR, and spectrograph systems, as well as other diagnostics clustered for a given set of geophysical research objectives.

The CEDAR Imager became operational on 1 September 1989 and is located on the Haystack Observatory complex in Westford, MA (USA). It is used in conjunction with the Millstone Hill radar and optical and radio facilities in the area. Typical results to be expected from such a cluster of instruments observing geophysical effects are treated in Mendillo et al., (1987). The imaging system itself may be readily applied to active experiments, as shown in Baumgardner and Karandanis (1984) for an earlier generation system.

Figure 4 gives a schematic of the CEDAR imager. When operated in the all-sky mode, light from 2π steradians is imaged onto the field lens where it is re-directed into the collimator. After the collimator, the light is in parallel bundles 100mm in diameter with a maximum angle of any ray to the optical axis of 4.5°. This small angle allows the use of interference filters with band passes of ~ 12Å HPFW. After passing through the filter the light is re-imaged at f/1 onto the photocathode of the intensified CCD. This detector is housed in a refrigerated housing and is kept at near -40°C to reduce its dark signal and thus allow the camera to integrate for 1 minute before the dark signal becomes significant. Figure 5 shows the sensitivity curves for the CEDAR detector and Figure 6 shows the performance of the detector under typical nightglow conditions.

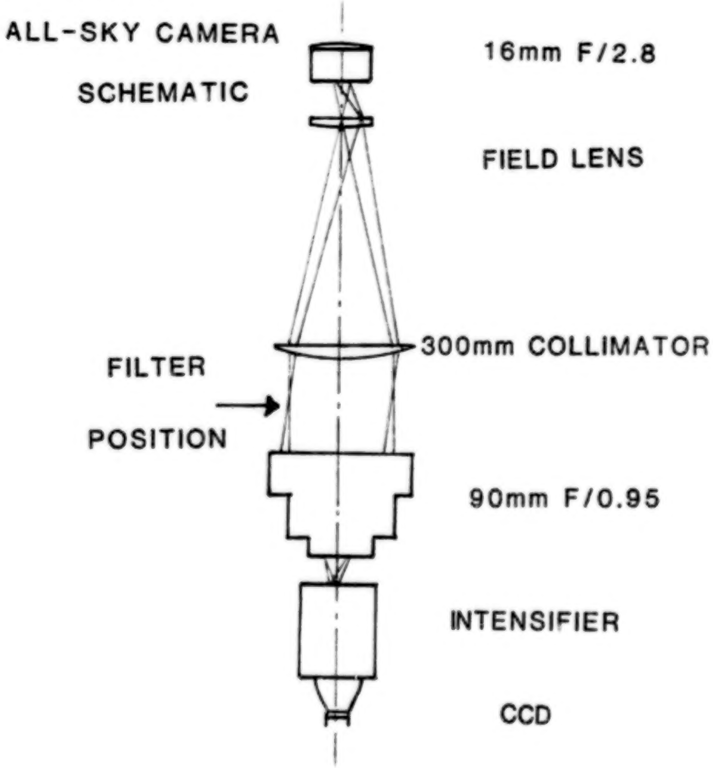


Figure 4. Schematic Diagram of CEDAR Imager.

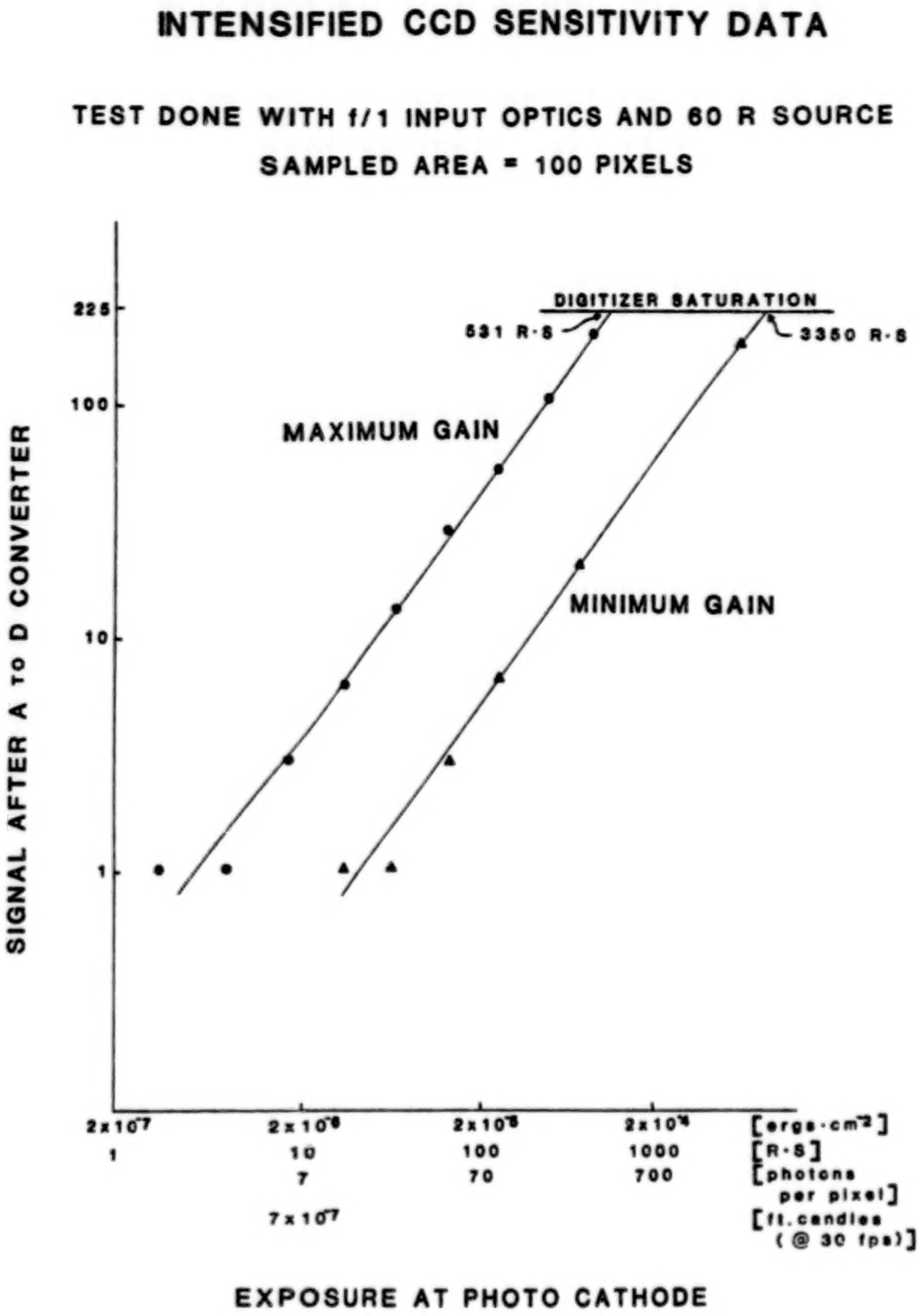


Figure 5. Sensitivity curves for CEDAR Detector.

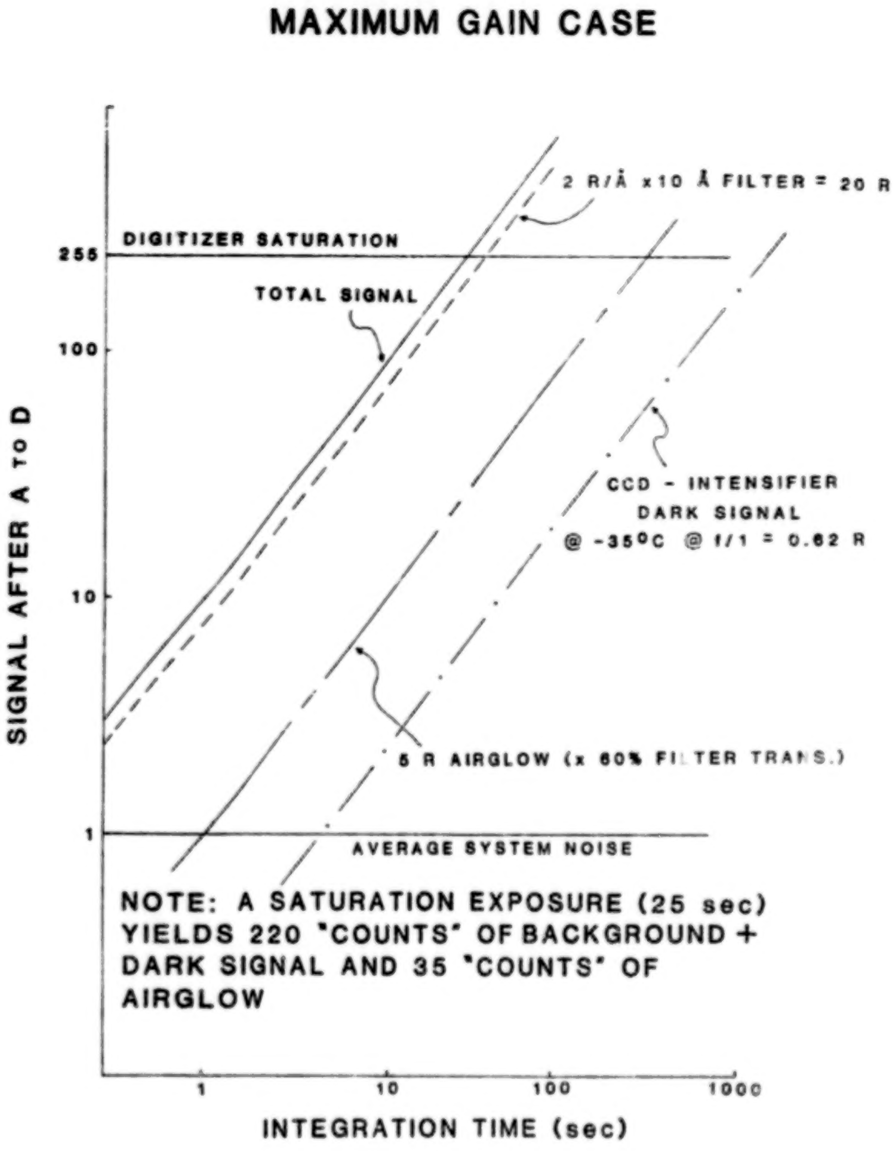
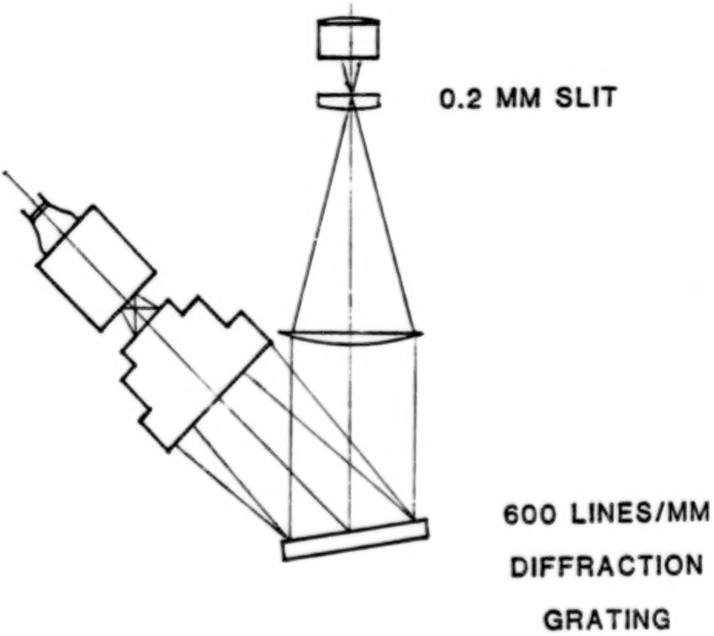


Figure 6. Performance curves of CEDAR detector under typical nightglow conditions.

III.3. Meridional Imaging Spectrograph

In many active experiments there are reasons to search for optical emissions over extended spatial and spectral domains. Investigations of geophysical structures (e.g., aurora) often also require simultaneous multi-wavelength coverage across extended features. To address these needs, a new meridional spectrograph was designed and tested (Baumgardner, 1989). Figure 7 gives a schematic of the prototype system. Its design is very similar to the CEDAR imager except that a reflection grating is used in place of the interference filter, and a slit is placed on the plano surface of the field lens. At the detector, the image of the slit at each wavelength has 360 spatial resolution elements corresponding to 180° field of view of the all-sky lens. The projected slit width is $\sim 1°$. In the dispersion directon there are 488 pixels, with the instrumental profile covering $\sim$ 10Å and a spectral range of 2000Å for any given grating position.



- Spectral Resolution ≈ 10Å
- Wavelength range/image ≈ 2000Å
- Angular resolution in all sky mode ≈ 1°
- Sensitivity @ 6300 with S/N ≥ 1 ≈ 20R•S

Figure 7. Schematic diagram of Meridional Spectrograph.

Figure 8, gives an example of a wavelength slice through a 30 second integration of nightglow from ~ 6300Å to 8300Å. The photocathode wavelength sensitivity curve has not been removed and is responsible for the downward slope of the spectrum toward the red end.

Since each integration can be saved on a WORM (Write Once Read Many) optical disk drive and each spectra contains spatial information along the slit, an image for each wavelength can be reconstructed showing space, time and intensity. Through a different choice of front end optics, the field of view can also be reduced from all-sky (meridional) to a narrow field mode more appropriate for small scale active experiments.

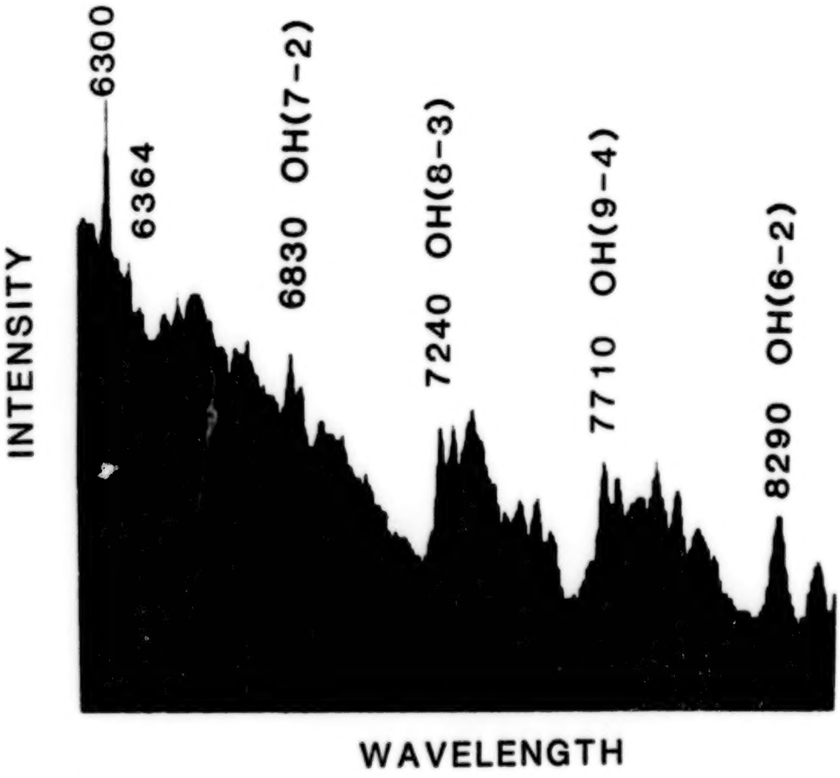


Figure 8. Zenith slice of spectrum taken with Meridional Spectrograph.

As a final point, it should be emphasized that the acquisition of digital images using CCD technology leads to a scale of data analysis work not encountered with earlier film-based systems. The advantages of working with digital images outweigh their drawbacks, but significant resources (personnel, hardware and software) are required to extract the full scientific yield from such systems. In Figure 9, we present in schematic form the data acquisition and processing system required for timely reduction and analysis of both CEDAR imager and meridional spectrograph images.

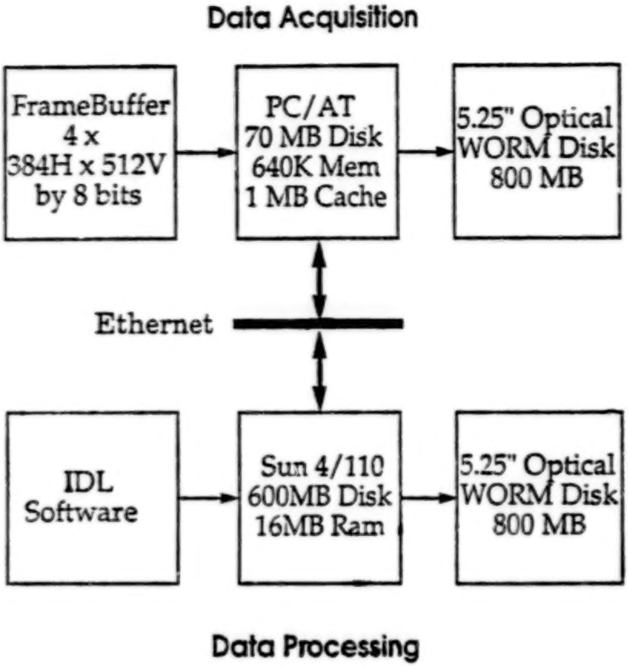


Figure 9. Data acquisition and processing schematic for CEDAR-type detector.

Acknowledgements. This work was supported, in part, by NASA grants, # NAGW-1629, and ATM# NAS8-36324, and by NSF grant, # ATM-8709487. We are grateful to Professor William Oliver for suggestions and discussions covering the incoherent scatter radar material, and to Ms. Kathryn Collins for her careful preparation of the typescript.

References

Baumgardner, J., and S. Karandanis, CCD System Using Video Graphics Controller, *Electronic Imaging, 3*, 28, 1984

Baumgardner, J., A Meridional Spectrograph for Aeronomic Studies, *EOS, 70*, 404, 1989.

Bernhardt, P.A., A Critical Comparison of Ionospheric Deplection Chemicals, *J. Geophys. Res., 92*, 4617, 1987a.

Bernhardt, P.A., Roussel-Dupre, R.A., Pongratz. M.B., Haerendel, G., Valenzuela, A., Gurnett, D.A., and R.R. Anderson, Observations and Theory of the AMPTE Magnetotail Barium Releases, *J. Geophys, Res., 92*, 5777, 1987b.

Bernhardt, P.A., Tepley, C.A., and L.M. Duncan, Airglow Enhancements Associated with Plasma Cavities Formed During Ionospheric Heating Experiments, *J. Geophys. Res., 94*, 9071, 1989.

Bernhardt, P.A., Kashiwa, B.A., Tepley, C.A., and S.T. Noble, Spacelab-2 Upper Atmospheric Modification Experiment Over Arecibo, 1, Neutral Gas Dynamics, *Astrophys. Lett. Comm., 27*, 169, 1988.

Brenning, N. (editor), Critical Ionization Velocity Phenomena, *Adv. Space Res., 10,* (8), 1989.

Burke, W.R. (compiler), Active Experiments in Space, *Proceedings of Symposium at Alpback*, Austria, 24-28 May, 1983, European Space Agency Report ESA-SP-195, Paris, France.

CEDAR, Coupling, Energetics and Dynamics of Atmospheric Regions, Vol. 1. Overview, Aeronomy Program, Nat. Sci. Found., Washington, DC, USA 1986.

Evans J.V., High-Power Radar Studies of the Ionosphere. *Proc. IEEE, 63,* 1636, 1975.

Evans J.V., Theory and Practice of Ionosphere Study by Thomson Scatter Radar. *Proc. IEEE, 57,* 496, 1969.

Foppl, H., Haerendel, G. Haser, L., Loidl, J., Lutjens, P., Lust, R., Meltzner, F., Mayer, B., Neuss, H., and E.R. Rieger, Artificial Strontium and Barium Clouds in the Upper Atmosphere, *Planet. Space Sci., 15*, 357, 1967

Haerendel, G. and M. Mendillo (editors), Active Experiments, *Adv. Space Res., 8*, (1), 1988.

Haerendel, G., Modification of Ionized Media by Chemical Substances — A Review of Physical Processes, AGARD Conference Proceedings No. 192, (H.J. Albrecht, ed.) 13-1, AGARD (Bruxelles), 1976

Heppner, J.P., Stolarik, J.D., and E.M. Wescott, Electric Field Measurements and the Identification of Currents Causing Magnetic Disturbances in the Polar Cap, *J. Geophys. Res., 76*, 6028, 1971.

Holt J.M., Rhoda D.A., Tetenbaum D. and VanEyken A.P., A New Technique for the analysis of Incoherent Scatter Lag Profile Measurements. Presentation at COSPAR XXVII, Helsinki, 1988.

Mende, S.B., Swenson, G.R., Geller, S.P., Doolittle, J.H., Harendel, G., Valenzuela, A. and O.H. Bauer, Dynamics of a Barium Release in the Magnetosphere Tail, *J. Geophys. Res.,* (in press), 1989.

Mende. S.B. and S.E. Harris, Measure of the Line of Sight Velocity of High Altitude Barium Clouds: A Technique, *Applied Optics, 21*, 3348, 1982.

Mendillo, M. (editor), Active Experiments in Space, *Adv. Space Res., 10*, (8), 1989.

Mendillo, M., Ionospheric Holes: A Review of Theory and Recent Observations, *Adv. Space Res., 8*, 51, 1988.

Mendillo, M., and J. Baumgardner, Optical Signature of an Ionospheric Hole, *Geophys. Res. Lett., 9*, 215, 1982.

Mendillo, M., and J. Baumgardner, Airglow Characteristics of Equatorial Plasma Depletions, *J. Geophys. Res., 87*, 7641, 1982b.

Mendillo, M., Baumgardner, J., Aarons, J., Foster, J., and J. Klobuchar, Coordinated Optical and Radio Studies of Ionospheric Disturbances: Initial Results from Millstone Hill, *Annales Geophysicae, 5A*, 543, 1987.

Mendillo M., Baumgardner J., Allen D.P., Foster J., Holt J., Ellis G.R.A., Klekociuk A., Reber G., Spacelab-2 Plasma Depletion Experiments for Ionospheric and Radio Astronomical Studies. *Science, 238,* 1260, 1987.

Mendillo M. and Forbes J., Theory and Observation of a Dynamically Evolving Negative Ion Plasma. *J. Geophys. Res., 87,* 8273, 1982.

Moorcroft D.R., On the Determination of Temperature and Ionic Composition by Electron Backscattering from the Ionosphere and Magnetosphere. *J. Geophys. Res., 69,* 955, 1964.

Rees, D., Greenaway, A.H., Gordon, R., McWhirter, I., Charleton, P.J. and A. Steen, The Doppler Imaging System: Initial Observations of the Auroral Thermosphere, *Planet. Space Sci., 32*, 273, 1984.

Rees, D., Hallinan, T.J., Stenback-Nielsen, H.C., Mendillo, M., and J. Baumgardner, Optical Observations of the AMPTE Artificial Comet From the Northern Hemisphere, *Nature, 320,* 704, 1986.

Rosenbluth M.N., Rostoker N., Scattering of Electromagnetic Waves by a Non-Equilibrium Plasma. *Phys. Fluids, 5,* 776, 1962.

Stenbaek-Nielsen, H.C., Hallinan, T.J., and E.M. Wescott, Acceleration of Barium Ions Near 8000 km Above an Aurora, *J. Geophys. Res., 84*, 10788, 1984

Sultan P.J., Mendillo M., Oliver W.L., Holt J., Detecting Heavy Negative Ions With an Incoherent Scatter Radar During Active Experiments, *EOS*, Nov. 1989.

Wand R.H. and Mendillo M., Incoherent Scatter Observations of an Artificially Modified Ionosphere. *J. Geophys. Res., 89,* 203, 1984.

Wescott, E.M., Stenbaek-Nielsen, H.C., Hallinan, T., Foppl, H. and A. Valenzuela, Star of Condor: A Strontium Critical Velocity Experiment, Peru, 1983, *J. Geophys. Res., 91*, 9933, 1986.

Chapter 10

MODIFICATION OF THE IONOSPHERE AS A METHOD OF ITS DIAGNOSTICS

L. M. Erukhimov and N. A. Mityakov

Radiophysical Research Institute, Gorky, USSR

Among many of the problems of ionospheric physics there are ones associated with the definition of characteristic constants and the hierarchy of processes defining ionospheric properties under different geophysical conditions. In traditionally passive experiments their solution meets with sufficiently large difficulties. The aim of the present paper is to discuss briefly a range of problems of ionosphere diagnostics which can be solved using modern techniques and methods of ionosphere modification. The question deals with the use of weak ionospheric plasma disturbances, created by radio wave beams, to study its dynamics and define a number of its parameters. Here the basic differences of the methods based on the ionosphere modification from passive ones is in the possibility of regulation of the processes occurring and creation of controlled initial conditions. Briefly, the modification of the ionosphere by a powerful radio wave field is based on the following processes [*Ginzburg et al.*, 1979; *Gurevich and Shvartsburg*, 1973; *Gershman et al.*, 1984; *Grach et al.*, 1989]. An electron (with a charge e and mass m) in the radio wave field $E = E_\omega e^{i(\omega t - \mathbf{kr})}$ (for simplicity the wave frequency ω is much larger than the frequency of electron collisions with other particles ν_e and its gyrofrequency ω_H) receives the velocity

$$\mathbf{V}_\omega = i\,\frac{e\mathbf{E}_\omega}{m\omega}$$

in the result of which the function of the electron velocity distribution is changed $f(\mathrm{v}) \to f(\mathrm{v} + \mathrm{v}_\omega)$. Here phenomena of hydrodynamic character occur when the variations of the distribution function $f(\mathrm{v})$ as a whole are important, as well as the kinetic effects where variations of $f(\mathrm{v})$ only in the definite region of velocities play the main role. The role of hydrodynamic effects is convenient to show by an example of equations for the average velocity v, concentration N and temperature T of electrons

$$\frac{\partial \mathbf{v}}{\partial t} = \Sigma\, \nu_{e\beta}(T)(\mathbf{V} - \mathbf{V}_\beta) - (\mathbf{v}\nabla)\,\mathbf{v} - \frac{e}{mc}\left[\mathbf{v}_\omega\, \mathbf{H}_\omega\right] - \frac{1}{Nm}\nabla NT \tag{1}$$

$$\frac{\partial T}{\partial t} = \frac{2}{3}\,\sigma'_\omega\, E_\omega E^*_\omega - \delta_{e\beta}\,\nu_{e\beta}\,(T - T_\beta) + \chi\Delta T \tag{2}$$

$$\frac{\partial N}{\partial t} + div\, N_\omega \mathbf{v}_\omega = q(T) - \alpha^1(T)\, N_e^2 \tag{3}$$

Here $\nu_{e\beta}$ is the frequency of electron collisions with other particles, $\mathbf{H}_\omega$ is the strength of the wave magnetic field, σ'_ω is the high frequency specific conductivity ($\sigma'_\omega = \frac{e^2\nu_e}{m\omega}$, $\nu_e = \Sigma\,\nu_{e\beta}$), $\delta_{e\beta}$ is the portion of energy transmitted by electrons to other particles in collisions, χ is the specific heat conductivity, q and α^1 are the functions describing ionization and recombination of particles, $N_\omega = \frac{k\mathrm{v}_\omega}{\omega} N$ is high-frequency oscillations of the density in the longitudinal wave field.

From (1)-(3) it is seen that nonlinearity manifests itself in all three equations. In a strongly collisional plasma the temperature effects play the main role: the occurrence of a nonlinear addition in the friction force, resulting in variation of the current in plasma, change of the motion in plasma due to the pressure force $F_T = \nabla NT$ of the gas heated by radio waves (see (2)), variation of the effective coefficients of electron disappearance (their recombination and attachment, etc.) which are described by (3). Collisionless processes are due to the force of the light pressure (striction force F_S) and in this case Equation (1) describes scattering and diffusion of the gas due to the light pressure and processes associated with the nonlinear component of the electron current (3). In all cases the variations in the ion plasma component are due to polarization fields induced by electrons.

Collisionless plasma escaping from the region of the strong field along the force lines of the geomagnetic field $\mathbf{h} = \frac{\mathbf{H}_0}{H_0}$ becomes essential if the scale of the field inhomogeneity L_E is smaller than the electron free path $l_e = \frac{v_{Te}}{\nu_e}$ (v_{Te} is the thermal velocity of an electron). If $l_e \ll L_{E\parallel}$, then at "large" times $t > (\delta \nu_e)^{-1}$ always $F_T > F_S$. When there is inhomogeneity of the field $L_{E\perp}$ in the direction orthogonal to $\mathbf{h}$, then due to small transverse heat conductivity of the plasma, F_S is always smaller than F_T, if $L_{e\perp} > \rho_e$ ($\rho_e = v_{Te}/\omega_H$ is the gyroradius of an electron, ω_H is the gyrofrequency). Really, from (2) it follows (time t is read from the moment of the field switching on):

$$T = \frac{2}{3}\,\frac{e^2}{m\omega^2\delta_1}\left(1 - e^{-\delta_1 \nu_e t}\right) E_0^2,$$

$$\delta_1 = \delta + \frac{\rho_e^2}{L_{E\perp}^2} + \frac{l_e^2}{L_{E\parallel}^2}$$

and

$$F_T = \begin{cases} \dfrac{F_S}{\delta_1 \nu_e} &, \quad t > (\delta_1 \nu_e)^{-1} \\ F_S &, \quad t < (\delta_1 \nu_e)^{-1} \end{cases}$$

where $F_S = \frac{\varepsilon - 1}{8\pi} \nabla |\mathbf{E}|^2 \simeq \frac{e^2 N}{2m\omega^2} \nabla |\mathbf{E}|^2$ is the striction force, and **E** is summarized from interacting fields of all modes among which there is not only the radio wave but the plasma mode induced by it. Chemical processes are important if their characteristic time τ_{ch} is smaller than the time of the plasma ejection from the region of inhomogeneous field

$$\tau_{ch} < L_E^2 / D_i\,, \quad D_i \simeq v_{Ti}^2 / \nu_{in}$$

(v_{Ti} is the thermal velocity of ions, ν_{in} is the frequency of their collisions with neutrals).

The minimal scales L_E are evidently defined by the wavelength of the radiation inducing the nonlinearity. A standing electromagnetic wave which is formed by interfering incident and reflected from the ionosphere radio waves (Figure 1) [*Grach et al.*, 1989] can serve as an example. Its characteristic field scale $L_E = \frac{\lambda}{2}$. For the radio wave frequency $f = 6$ MHz, the value $\lambda/2 = 25$ m. In the lower ionosphere ($Z = 50$–70 km) the frequency ν_{in} is large, so, $\tau_{ch} \ll \tau_D$ and variation of free electron density due to heating by a powerful radio wave is controlled by the temperature dependence of the speed of electron attachment to oxygen molecules. The rate of this process varies from 15 s^{-1} at $Z = 55$ km up to 0.1 s^{-1} at $Z = 75$ km [*Belikovich et al.*, 1986]. Thus, the standing structure of the formed wave intensity induces the spatial periodic stratification of free electron density. At large altitudes where the time of the diffusional plasma ejection (due to the force F_T) becomes smaller than τ_{ch}, the formation of artificial periodic irregularities (API) occurs due to plasma ejection from the regions of the standing wave antinooks where the electron temperature is larger. Demolition of the plasma lattice after switching off the powerful radio transmitter creating the standing wave is explained by the same process of the molecular (and turbulent in the height region of 70-95 km) diffusion.

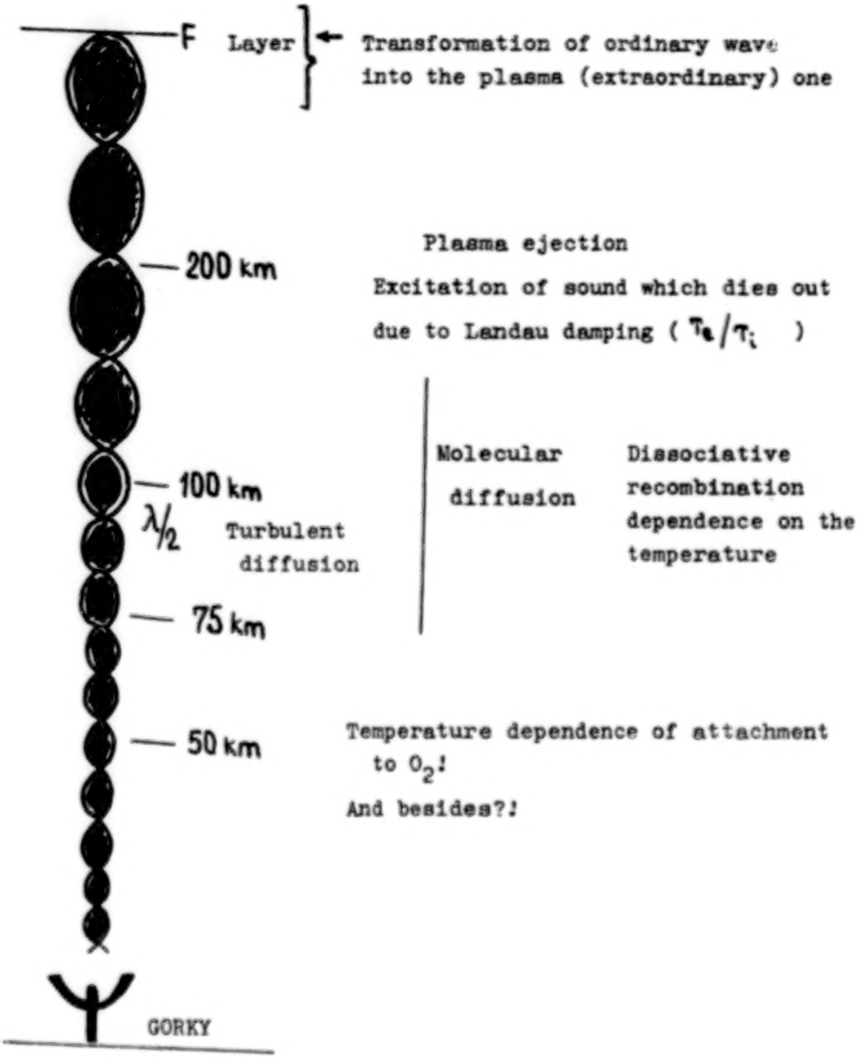

F Layer
Transformation of ordinary wave into the plasma (extraordinary) one
200 km
Plasma ejection
Excitation of sound which dies out due to Landau damping (T_e/T_i)
Molecular diffusion
Dissociative recombination dependence on the temperature
100 km
λ/2
Turbulent diffusion
75 km
50 km
Temperature dependence of attachment to O_2!
And besides?!
GORKY

Figure 1. Nonlinear effects in the interference of incident and reflected radio waves.

At last, at altitudes where the wavelength of the standing structure becomes smaller than the free path, the API occurrence and relaxation is controlled by ejection processes due to the striction force. Here the ion acoustic waves are excited which are damped owing to collisionless Landau mechanism of thermal ions. The rate of IAW damping depends on how their phase velocity $\left(v_{ph} = \sqrt{\frac{T_e+T_i}{M_i}}\right)$ is close to the ion thermal velocity $\left(v_{Ti} = \sqrt{\frac{T_i}{M_i}}\right)$ where the damping is maximal. In other words, the rate of damping depends on the relation T_e/T_i.

Figure 1 shows a possibility of one of the methods of ionospheric parameter study — scattering of probe radio waves by artificial periodic irregularities which is described below [*Gershman et al.*, 1984; *Grach et al.*, 1989; *Belikovich et al.*, 1975, 1977, 1986]. If in the interaction with plasma besides the radio wave, one more wave takes part with the wavelength λ_l smaller than the wave λ, then the scale L_E is defined by the length λ_l. This scale is minimal in the case of longitudinal plasma waves ($\lambda_l > r_{De}$, $r_{De} = \frac{v_{Te}}{\omega_{oe}}$ is Debye length, $\omega_{oe} = \left(\frac{4\pi e^2 N}{m}\right)^{1/2}$ is the Langmuir frequency). In the ionosphere λ_l can accept values from one centimeter up to dozens of meters. Excitation of plasma waves in the ionosphere by ground-based sources is realized because of the process of the wave transformation. Figure 2 illustrates schematically the dispersion curves of high-frequency ordinary and extraordinary waves in the plane "refraction index square n^2 — the relation square between plasma and operation frequencies" [*Ginzburg et al.*, 1979; *Gurevich and Shvartsburg*, 1973; *Gershman et al.*, 1984; *Grach et al.*, 1989; *Ginzburg*, 1967].

It is seen that a radio wave of extraordinary polarization sent vertically upwards to the ionosphere is reflected from its level of reflection out of the reach of the region where plasma waves exist. However, radio waves of ordinary polarization in the interval of altitudes from Z_1, where $\omega = \sqrt{\omega_{oe}^2 + \omega_{H}^2}$ up to Z_2 ($\omega = \omega_{oe}$) are present in this region. Thus, a transition can occur of a radio wave into a slower ($n > 1$) plasma one. Such transition is just called the mode transformation. It can occur by the diffraction way when dispersion curves draw closer to one another due to large-scale gradients in the medium, and it can take place due to the medium inhomogeneities (quasi-stationary in the given case) being polarized by the radio wave (scattering in plasma mode).

The plasma waves formed can be diagnosticated both by the method of incoherent scattering and by the artificial ionospheric radiation (AIR) which is the consequence of their back transformation into the radio wave. With a sufficiently large amplitude E_l which is equal to its maximal possible value

$$E_l = \delta N \frac{\omega}{\nu_e} E_0 >> E_0$$

in a homogeneous medium they induce nonlinear forces F_S and F_T and form small-scale plasma irregularities with $l \sim \lambda_l/2$. These irregularities are created more easily when they are elongated along the geomagnetic field force lines, i.e., they are stratified in transversal to **h** direction, as in this case the processes of fast plasma transfer along **h** do not have fatal consequences. In this case they are just the low-frequency anisotropic plasma turbulence which induces the intensive aspect scattering of radio waves of decameter-meter ranges. These irregularities amplify the radio wave transformation into the plasma one and vice versa, making this process strictly nonlinear and the plasma state unstable. Excitation of plasma waves and low-frequency disturbances due to nonlinear transformation is possible in the case of the ionosphere modification by two high-frequency waves with a difference between frequencies equal, for example, to the frequency ω_l of the Langmuir wave.

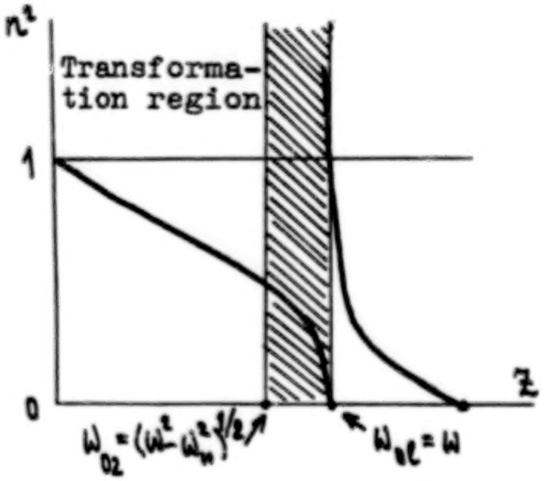


Figure 2. The region of radio wave transformation into plasma ones in the ionosphere.

Plasma waves are also subjected to decomposition processes. An example can serve their decomposition into a plasma wave of smaller frequency and a low-frequency wave, ion-acoustic, for instance

$$\omega_l = \omega_l' + \Omega_S ,$$

or a quasi-transversal helicon with $\omega \gtrsim \sqrt{\omega_H \Omega_H}$ (Ω_H is the gyrofrequency of ions).

In particular, frequencies of the artificial ionospheric radiation smaller than those of pumping waves are usually associated with the decomposition processes. Radiation with frequencies $\omega_{\mathrm{AIR}} > \omega_{PW}$ can be induced by processes of low-frequency and plasma wave coalescence, the vertical transfer of waves by accelerated electrons, particle radiation in fluctuating alternating electric fields, etc. (see Figures 3 and 4), in particular, with four-wave interaction.

Plasma waves induce particle acceleration which can transfer them not only in the vertical direction but also, by exciting the atmospheric components, induce the optical airglow of the atmosphere, as well as the artificial atmosphere ionization, at altitudes of the F-layer which are insufficient so far [*Grach et al.*, 1989].

Above we consider processes of spatial plasma stratification in the fields of powerful radio waves under the action of forces F_S and F_T giving the minimal scales of disturbance. There are a number of nonlinear processes (in particular, self-focusing instabilities) which lead to spatial stratifications of larger scales [*Gurevich and Shvartsburg*, 1973]. Processes not associated with the action of these forces are the above-mentioned chemical processes as well as those caused by variation of collision frequency ν_e. First of all these are the well-known effects of self-action and cross-modulation [*Gurevich and Shvartsburg*, 1973; *Gershman et al.*, 1984] and the effect of the generation of low-frequency ionospheric radiation-Getmatsev effect [*Gershman et al.*, 1984; *Grach et al.*, 1989; *Belyaev et al.*, 1987].

The sense of the Getmatsev effect is simple: by modulating the electron current structure due to the periodic variation of the friction force (electron conductivity), we can obtain alternating electric current which must radiate radio waves at the modulation frequency. The radiation received on the Earth is associated not only with the current value and its spatial structure but also with the conditions of low frequency Ω radio waves escaping from the ionosphere, which are defined by the depth of skin-layer of the ionosphere for the frequency Ω. By making measurements of the value and polarization of artificial low-frequency radiation at different frequencies one can investigate the radiation structure, the current structure at 70-80 km heights and hence, the parameters depending on them: the neutral gas motion speed $\mathbf{v}_n$ at midlatitudes and electric fields $\mathbf{E}_\perp$ in high latitudes. Variation of electron velocity transverse to geomagnetic field $\mathbf{h}$ due to the friction force by the neutral gas moving orthogonally to $\mathbf{h}$ with a speed $\mathbf{v}_n$ is equal to [*Gershman et al.*, 1984]

$$\mathbf{v}_1 = \frac{\nu_e^2}{\omega_H^2 + \nu_e^2}\,\mathbf{v}_n + \frac{\nu_e\,\omega_H}{\omega_H^2 + \nu_e^2}\,[\mathbf{v}_n\,\mathbf{h}]$$

The presence of the electric field $\mathbf{E}_\perp$ leads to electron motion with a speed

$$\mathbf{v}_2 = \frac{\nu_e}{\omega_H^2 + \nu_e^2}\,\frac{e}{m}\,\mathbf{E}_\perp + \frac{\omega_H}{\omega_H^2 + \nu_e^2}\,\frac{e}{m}\,[\mathbf{E}_\perp\,\mathbf{h}]$$

In the simplest case the low-frequency current component $j_\Omega = eN\mathrm{v}_\Omega \propto \dfrac{\Delta\,\nu_e\,\Omega}{\nu_e} = \dfrac{5}{6}\,\dfrac{\Delta\,T_\Omega}{T}$ where $\Delta\,T_\Omega$ is the amplitude of periodic variations of the electron temperature under the action of a modulated radio wave. Here variations of an electron speed v_Ω are maximal in the region where $\nu_e \sim \omega_H$. The modification of the lower ionosphere becomes more effective in operation with a high power as well as with approximation of the action frequency to the electron gyrofrequency. In this

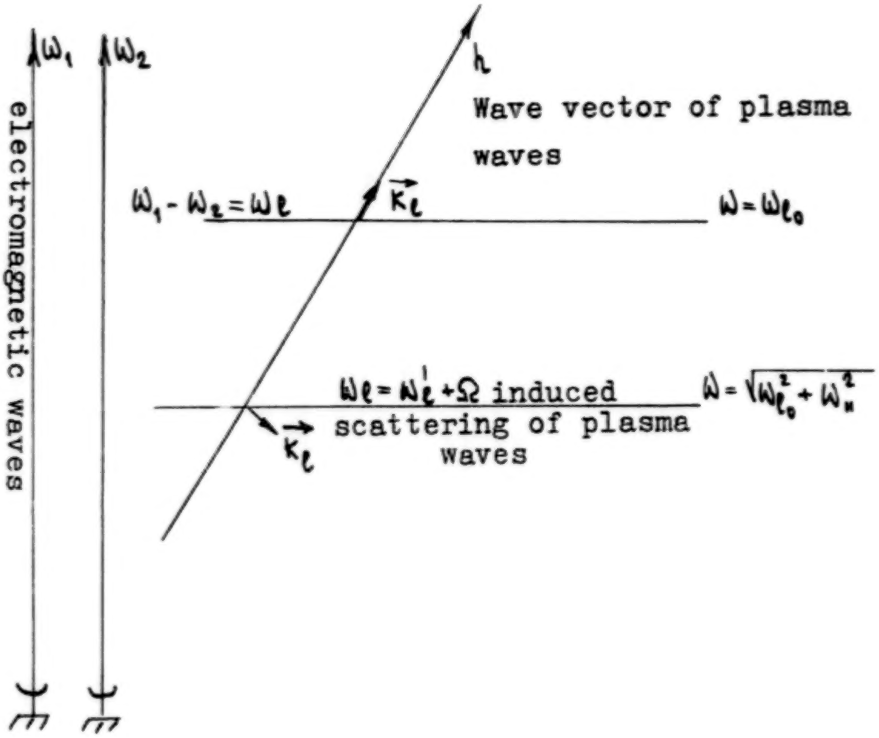

ω₁
ω₂
electromagnetic waves
h
Wave vector of plasma waves
ω₁ - ω₂ = ωl
kl
ω = ωl0
ωl = ω'l + Ω induced scattering of plasma waves
ω = √(ωl0² + ωH²)
kl

Figure 3. Generation of plasma and ion-acoustic waves in ionosphere.

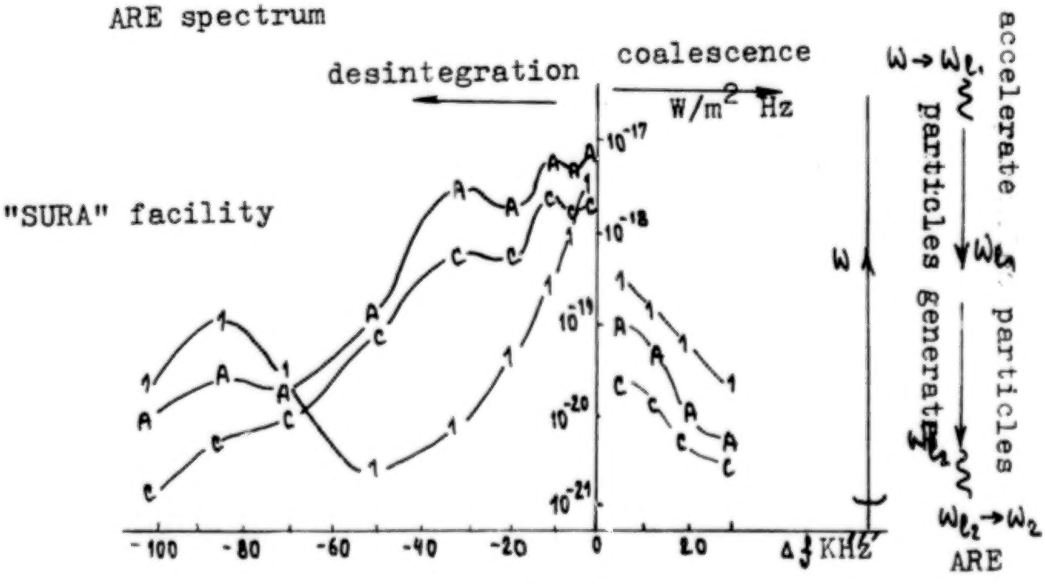


Figure 4.

connection a new stage began in the investigation of the Getmatsev effect in the eighties when in Tromsø (Norway) and Vasil'sursk powerful wide-band heating facilities were introduced [*Stubbe et al.*, 1985; *Belov et al.*, 1983]. A broadening of the frequency wave range and an increase of the heating facility power permit one to essentially increase the signal-to-noise ratio and obtain the stability for the combination frequency signal (CFS) reception in the frequency range of modulation current systems 0.5 kHz ÷ 10 Hz. This gives a possibility to investigate the fine structure and dynamics of the neutral component (at midlatitudes) as well as study a wider class of geophysical problems including methods of action on the magnetospheric plasma instabilities. Figure 5 presents CFS spectra at small displacements of the observation point over the horizontal from the current source (the dependence of the field magnetic component on the low-frequency radiation). From the figure it is seen that maxima are well distinguished in the near field which are associated with the fact that in the Earth-ionosphere waveguide there is an integer of half-waves of artificial low-frequency radiation. There is a shift of the first maximum from 2.2 kHz at midday to 1.6 kHz in the evening which corresponds to the diurnal variation of the upper waveguide wall height during the day (from 68 km in daytime up to 88 km at night). Note that the CFS method permits using the swinging of the powerful radiator directivity pattern (or simultaneous radiation in three directions) to measure over the well-known three-point method — in the given case from three CFS sources — the value and direction of the neutral wind velocity $\mathbf{v}_n$ in the horizontal plane of its variations. Information on vertical velocities is obtained in CFS reception at two different frequencies Ω_1 and Ω_2, since their sources are spaced over the height. In the case of one-point reception, the information on the velocity $\mathbf{v}_n$ variation is obtained from the polarizational variations of CFS.

One of the perspective directions for the Getmatsev effect development is connected with the generation of ultra-low-frequency electromagnetic fields $\left(\frac{\Omega}{2\pi} < 10 \text{ Hz}\right)$ (geomagnetic pulsations) especially in the frequency region corresponding to the frequencies of the ionospheric Alfven resonator [*Belyaev et al.*, 1987].

Further CFS investigations are expected to be related to the total exploitation of geomagnetic pulsation wave range (up to 10^{-3} Hz), with excitation of Alfven ionospheric and magnetospheric resonators by the modulation of their lower edge conductivity, especially in the region of auroral latitudes [*Belyaev et al.*, 1987; *Bespalov and Trakhtengerts*, 1986], that must result in strong variations of precipitating particle streams. Of great interest are CFS observations from the F region, in particular, under the condition of developed parametric instabilities.

Let us consider again the artificial periodic irregularities. Figure 6 shows the characteristic dependences of the relaxation time τ on altitude for three periods of observation. The variety of the given curves $\tau(Z)$ is striking both at small altitude, i.e., in the region of chemical processes dominating and in the altitude region where turbulent diffusion prevails and where the method can be the tool for its investigation. From this viewpoint a good illustration for the possibility of the vertical motion diagnostics serves Figure 7 where the time dependence is given for the amplitude of a signal scattered by the "lattice" in the periodic creation of it by impulses of 2 [illegible] duration. It is seen that in the repeated switching on of the powerful radio transmitter (at places indicated by a dotted semicircle) the rapid lattice breakdown occurs. The latter is associated with the fact that due to the vertical component of the turbulent and regular wind speed transporting the lattices, the breaking down of the preceding lattice occurs by "unphaseable" heating. On the basis of the given effects it is easy to develop a method for definition of the turbulent vortices of different scales as well as coefficients of the turbulent diffusion so necessary for constructing lower ionosphere models.

Figure 8 illustrates the altitude profiles of electron density obtained by the given method [*Belikovich et al.*, 1986]. The electron density N is defined from the requirements of spatial synchronism of the scattered wave and the artificial periodic irregularity — "lattice." It results in equality

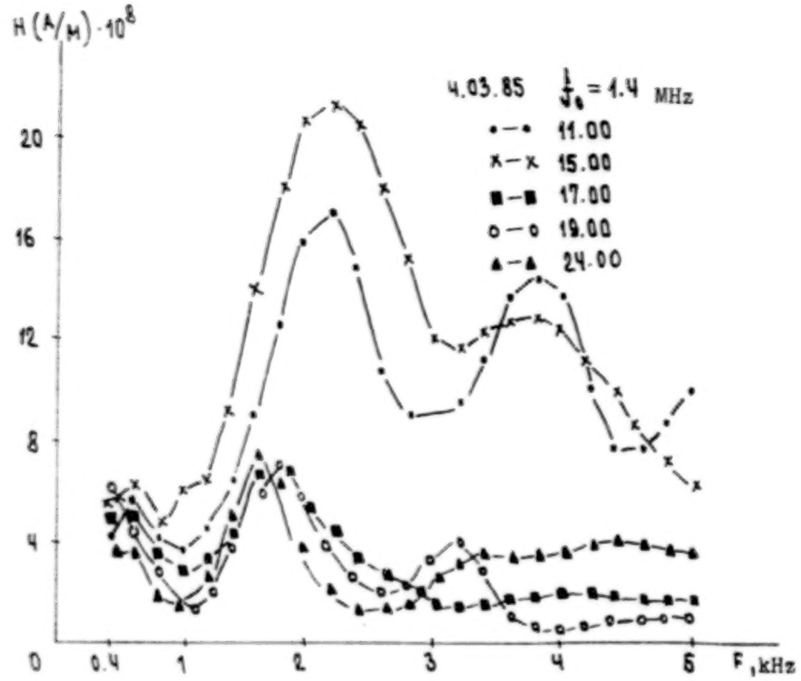


Figure 5. Spectra obtained by "SURA" facility in different times of day.

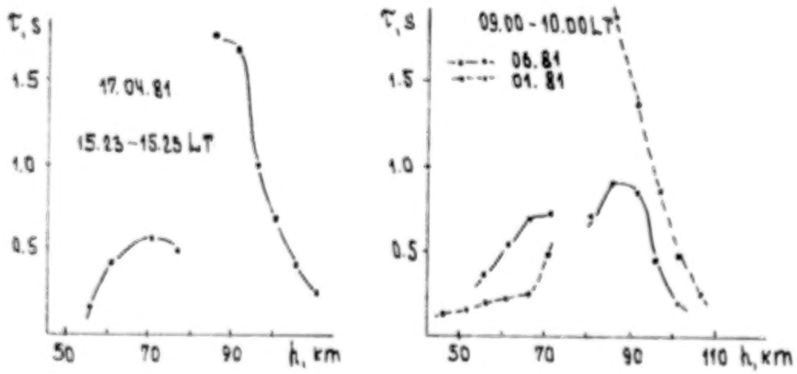


Figure 6. The dependence of relaxation time of API on altitude ("SURA" facility).

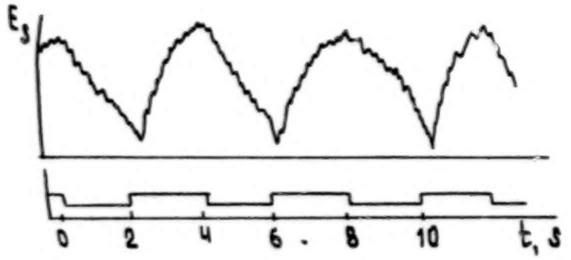


Figure 7. Breakdown of periodic irregularities at repeated switching of the SURA transmitter.

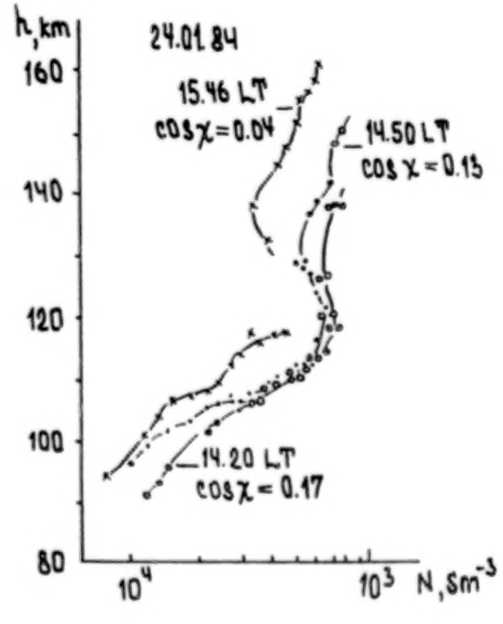


Figure 8. Altitude profiles N(h) obtained by the method of resonance scattering.

$$f_{PW}\, n^{\pm}\left(f_{PW}\right) = f_{Pr}\, n^{\mp}\left(f_{Pr}\right),$$

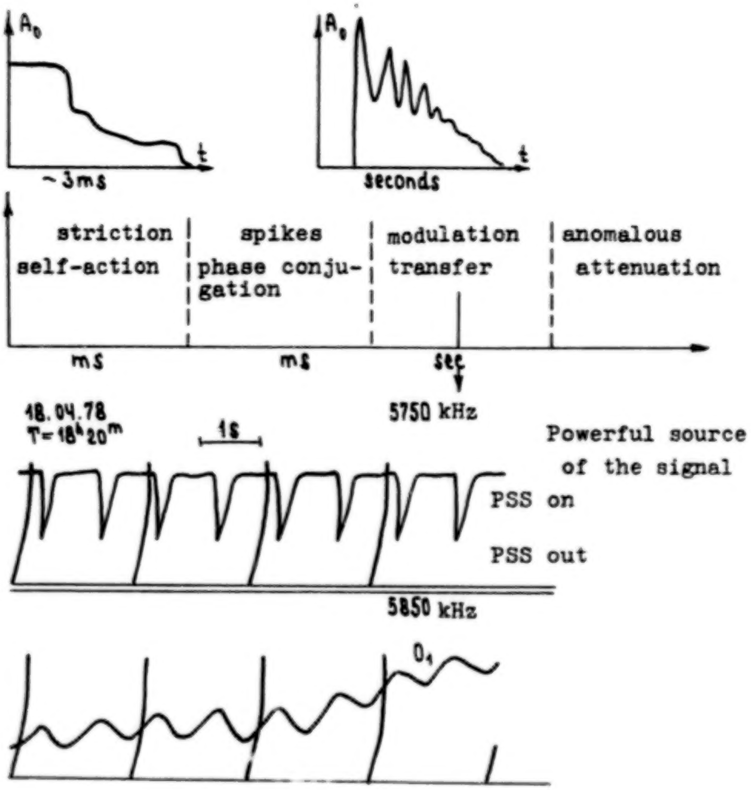
where $n^{\pm}\left(f_{PW}\right)$ and $n^{\mp}\left(f_{Pr}\right)$ are refraction indices of the medium for frequencies of the pumping wave $\left(f_{PW}\right)$ and the probe wave referred to different normal polarizations (+ – is the ordinary component). *N(Z)* is measured by changing the frequency of one of the transmitters and definition of the acting altitude of the backscattering over the strobed signal. API relaxation in the upper ionosphere is defined by the plasma ejection. The API formation in the F region of the ionosphere is accompanied by excitation of long wavelength quick damping ion-acoustic oscillations. For them the frequency $\Omega = k v_S$, where $v_S = \sqrt{\frac{T_e+T_i}{M_i}}$, and damping is determined by collisionless Landau damping, the decrement of which is defined by the ratio T_e/T_i (high altitudes) or collisional absorption determined by the frequency of collisions ν_{in} between ions and neutrals. This permits one to estimate the values T_e, T_i and ν_{in}.

In perspective, of interest is excitation of APIs which are elongated along the force lines **h**; they must be considerably more intensive due to small transfer coefficients in the direction orthogonal to **h**.

Let us discuss the experimental data on artificial ionospheric turbulence. The scenario of the development process of plasma turbulence from small to large times of development can be presented in the following way (Figure 9). At the exceeding of threshold values of the striction parametric instability development at times of the order of 1-3 ms, the anomalous attenuation of pumping waves due to striction excitation of plasma waves begins. This process is mainly developed near the level of PW reflection (Figure 2) due to the known effect of the field amplification near the reflection level. Then, since times of the development are inversely proportional to the square PW amplitude, the striction plasma wave excitation occurs at smaller altitude in the interval between the reflection level and the level of the upper hybrid resonance. Here, in the region of (UHR) the plasma waves are already orthogonal to **h**, and at the level of reflection $\mathbf{k}_l \parallel \mathbf{h}$ the excitation of plasma waves with $\mathbf{k}_l \perp \mathbf{h}$ (in the region of UHR) can be accompanied by the formation of low-frequency waves with $\mathbf{k}_\Omega$ approximately orthogonal to **h**.

The formation of low-frequency waves damping slower than the ion sound must lead to the decrease of excitation thresholds and correspondingly to larger times of the process development. It is natural to expect that initially the plasma waves must be excited in maxima of a standing electromagnetic field. However, even small shifts of the reflection point leads to a shift of the field maxima. The latter can be one of the reasons (others will be discussed below) of the breakdown of the effect of self-action at times of dozens and hundreds of milliseconds resulting in the observing oscillation regime (the effect of "spikes") of the pumping wave and probe one. Here, the slower the process develops, the larger interval of altitude it occupies and the more wide-band it is. If the PW is switched on and off periodically the effect of the modulation transfer is observed for probe waves at this intermediate phase. As Figure 9 shows, a phase shift can be observed (in time) of oscillations being transferred to probe waves. Even at this duration (dozens of milliseconds) the thermal effects begin to develop. The minimal time of their development is defined by ν_e^{-1} and the next over the rank is the time of longitudinal heat conductivity $\tau_D = \frac{l_{\parallel}^2}{4D_{e\parallel}}$, $D_{e\parallel} = \frac{T}{m\,\nu_e}$ (here T is in the energetic units), $l_{\parallel}$ is the scale of the temperature variations along **h**. The value τ_D amounts 0.05-0.2 s and depends on the longitudinal scale [*Erukhimov et al.*, 1987]. But it is convenient for irregularities to be developed self-consistently ejecting the ions in transverse to **h** directions, so that a considerable spatial charge cannot occur. The latter takes place if $\frac{l_{\parallel}^2}{D_{e\parallel}} \sim \frac{l_{\perp}^2}{D_{i\perp}}$, $D_{i\perp} = \frac{T_i\,\nu_{in}}{M_i\,\Omega_{H}^2}$. It

"Spikes" and the anomalous attenuation



The modulation depression effect

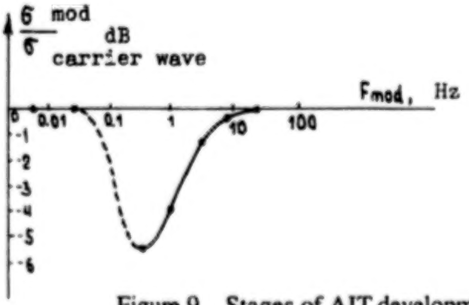


Figure 9. Stages of AIT development.

is not excluded that for this reason the time of "expectation" in the process of formation of thermal irregularities elongated along **h** increases with the growth of $l_\perp$.

Of interest is the scattered signal spectrum dynamics in the process of irregularity development. Initially, the narrow spectrum broadens gradually and that testifies to a nonlinear mechanism of formation of irregularity scale spectra at initial stages of artificial turbulence development due to interference of small-scale irregularities.

Then, larger irregularities stabilize the smaller ones at a lower level and the initially inverse spectrum $\phi_N(æ)$ of the turbulence which has a maximum in the region of small scales ($l_\perp < 1 \div 2$ m) becomes quasi-exponential as shown in Figure 10, where the spectrum $\phi_N(æ_\perp)$ is given, $æ_\perp$ is the transverse wave number $\left(æ_\perp = \frac{2\pi}{l_\perp}\right)$ obtained at the stationary phase by the method of the aspect scattering and the satellite signal scintillations.

Striction effects, "spike" effects, and that of modulation transfer vanish due to a stronger effect of the anomalous absorption of the pumping wave and plasma waves at upper hybrid resonance (UHR) levels. The energy of the pumping wave does not achieve the reflection level, practically completely damping at the UHR level. The period of the aspect scattering, scintillations, F_{spread} phenomenon, the occurrence of large-scale stratifications in the F layer, etc. take place.

After the powerful radio transmitter is switched off, the artificial ionospheric turbulence is relaxed; first small scales decrease and then larger ones decrease. Figure 11 shows the characteristic dependences of irregularity relaxation time τ_r of the stationary level of their transverse to **h** dimension. It is seen that there is a characteristic scale $l_\perp^*$ where there is a break in the character of the dependence $\tau_r(l_\perp)$: on $\tau_r \propto l_\perp^2$ for $l_\perp < l_\perp^*$ up to $\tau_r \propto l_\perp^{1/4}$ for $l_\perp > l_\perp^*$. For the second case the relaxation is defined by the longitudinal ion diffusion, and in the first case the diffusion coefficients are close to the transverse electron ones, $D_{e\perp} = \frac{2T\nu_e}{m\,\omega_{H}^2}$. It is not excluded that for $l_\perp < l_\perp^*$ the longitudinal scale of irregularities is proportional to $l_\perp$ and the relaxation proceeds with approximate "self-correlation" between the times of transverse electron and longitudinal ion diffusion. The relaxation is of two-step character. Its rate considerably decelerates after the intensity of artificial irregularities decreases by ten-fifty times (the upper part of Figure 11). The second step can be associated with the nonlinear pumpback of the turbulent energy by large scales. A problem on the mechanism of natural plasma turbulence formation is one of the keys in the physics of ionosphere and cosmic plasma physics. Since the artificial turbulence in the scale region $l_\perp < 1$–10 km is close in its properties to the natural one, its investigation serves for understanding the natural processes leading to the formation of different scales of irregularities and helps to solve key problems in the building of inhomogeneous structure models of the upper ionosphere. So, for example, a tremendous impression is made because of how easy F_{spread} is excited at low latitudes (Dushanbe), while under natural conditions it almost never appears.

It was noted long ago that irregularities caused by short-period action radio waves continue to develop after the powerful transmitter is switched off. The temperature fluctuations excited in the ionosphere, forgotten by their creator, form the plasma inhomogeneities which relax and reach the maximal value. But the time of the maximum and the velocity of inhomogeneity destruction depend on the pause between the successive radio wave switching, i.e., on the presence of inhomogeneities of other scales. This can also testify to the role of direct and inverse nonlinear pumping over the turbulent spectrum into the formation of definite scales of inhomogeneities and the characteristic lifetime. The dynamics of the frequency spectrum of the scattered signal in the turbulence relaxation is also informative.

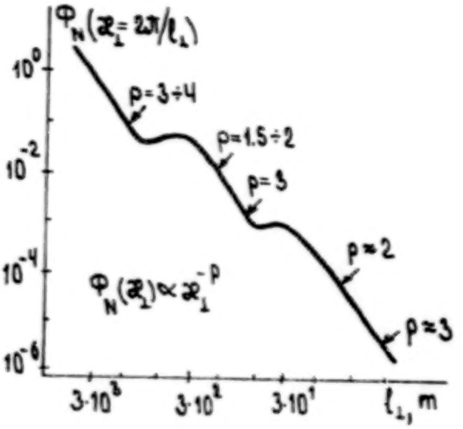
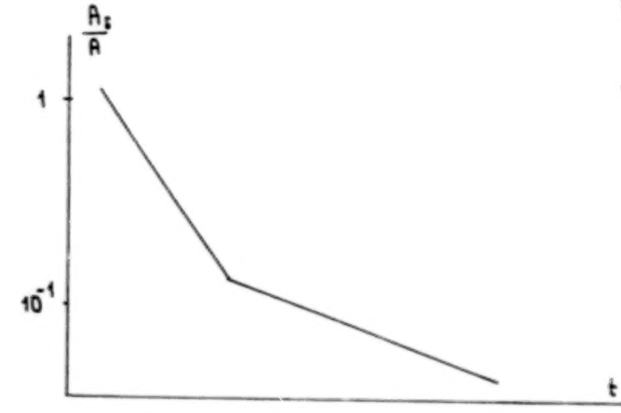


Figure 10. The spectrum of scales of artificial ionospheric turbulence.



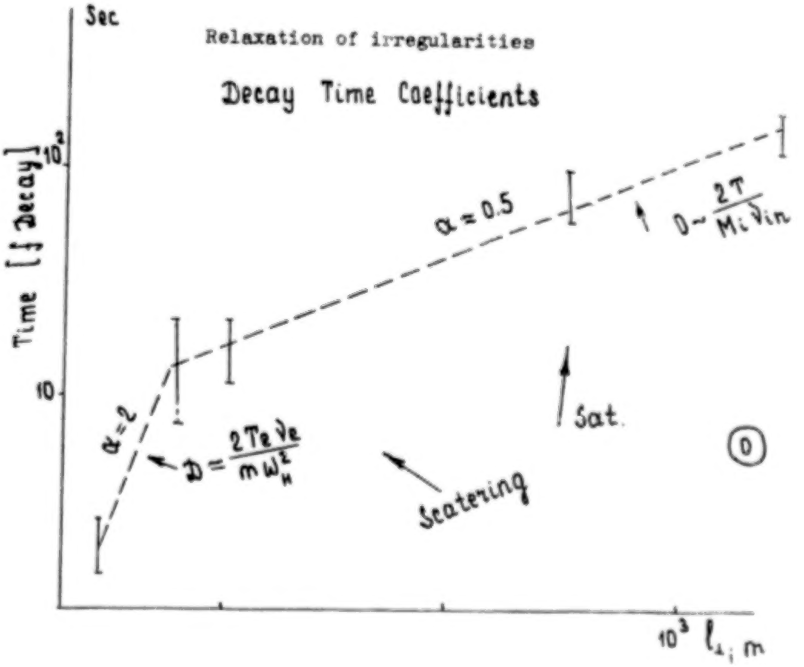


Figure 11. Illustration of two-step relaxation of the amplitude of the aspect signal and the dependence of the decay-time on the irregularity scale.

Of interest is the fact that, after the inhomogeneity relaxation, they can again occur in (~30 s), which is much larger than the time of their relaxation [*Belenov et al.*, 1977]. As shown in *Yampol'skij* [1989], this effect has a brightly expressed resonance nature, i.e., the ionosphere here serves as a memory device. Processes which are investigated in the ionosphere modification, especially when the question deals with the dynamics of small-scale plasma structures with the wave transformation, are of great interest in the cosmic plasma. Again, the atmosphere similarity parameters of a number of cosmic objects are close to those of the Earth. The solar chromosphere can serve as an example. Thus, at present, the ionosphere can already be presented as the most close to our cosmic laboratory, equipped with different diagnostic methods and possibilities of excitation of different types of instabilities.

Facilities of the Ionosphere Modification by Powerful Radio Waves

The heating nonlinearity is of importance for ionosphere modification. It is defined by the ratio of the wave field amplitude $E \sim 0.3\,R^{-1}\sqrt{PG}\ (V/m)$ to the characteristic plasma field E_P, where R is the distance from the source in km, P is the power of the transmitter in kW, G is the antenna amplification coefficient relative to the isotropic emitter. The nonlinear effects sharply increase when the wave frequency coincides with one of the plasma resonance frequencies — Langmuir ω_{oe}, gyrofrequency ω_H of electrons or with the frequency of the upper hybrid resonance $\sqrt{\omega_{oe}^2 + \omega_H^2}$. The duration of the ionospheric plasma modification effects for the heat nonlinearity amounts $\sim(\delta\nu)^{-1}$, where ν is the effective frequency of collisions of electrons and ions or molecules, and δ is the portion of energy losses by an electron in one collision. At last, the pumping wave polarization must correspond to the polarization of normal waves in the ionosphere (ordinary or extraordinary). All this permits one to formulate some general principles of building of investigation facilities for the modification of the ionosphere by radio waves. First of all, the facility must have the equivalent power $PG \gtrsim 10$ MW to provide the ratio $\frac{E^2}{E_P^2} \gtrsim 10^{-2}$ for the vertical radio wave beam in the interval of altitude 80–250 km. Since the characteristic time $(\delta\nu)^{-1}$ sharply increases with altitude, from 10^{-4} s up to 10 s, the duration of radiating radio pulses must change in the same limits. This means that the facility transmitter must operate both in the impulse regime and in the regime of continuous radiation. If the gyrofrequency $\frac{\omega_H}{2\pi} \simeq 1.4$ MHz is practically constant, the plasma frequency ω_{oe} (and hence, $\sqrt{\omega_{oe}^2 + \omega_H^2}$) is defined by the critical frequency of the ionosphere, diurnal and seasonal variations of which in midlatitudes are in the limits of 2–10 MHz. From here it is clear that the range of operating frequencies of the facility in the ideal case must approximately amount to 1.3–10 MHz. The limiting polarization of normal waves in the ionosphere in the vertical radio wave beam incidence is close to the circular one, practically at all latitudes apart from a narrow region of approximately ± 10° near the geomagnetic equator. So, the antenna must receive radio waves with right (or left) circular polarization.

The power of standard SW transmitters does not exceed 100–200 kW. To provide a sufficient ($\sim 10^2$) amplification coefficient of the antenna $G = \frac{4\pi S}{\lambda^2}$ the antenna systems are used with the maximally large area S. The potential of the facility can be increased if we use several (N) synchronously operating modula, each consisting of the transmitter loaded for a separate antenna section. The equivalent power of the facility in this case is equal to $PG = P_N\,G_N\,N^2$, where $P_N\,G_N$ is the equivalent power of each modulus. As the antenna, a horizontal phased antenna lattice is used. In the synphase lattice feed, the vertical directivity pattern is formed with a width $\Delta\theta \simeq \frac{\lambda}{D}$, where D is the dimension of the antenna field. By varying the separate lattice elements, one can control the directivity pattern in the limits ± 40° from the zenith. A separate element of the lattice is two horizontally crossed wide-band vibrators located at the height $\frac{\bar{\lambda}}{4}$ above the Earth ($\bar{\lambda}$ is the mean wavelength over the range). The crossed vibrators are fed with a phase shift $\pm\frac{\pi}{4}$, to provide the wave radiation with right (or left) circular polarization. The coefficient of the frequency overlapping of a wide-band vibrator does not exceed two magnitudes. So, for broadening of the frequency range the facilitiy uses two (or three) independent antenna systems.

All facilities for modification of the ionosphere by radio waves are built over the given scheme and differ by the frequency range and equivalent radiation power. The facility parameters are given in Table I where we give the name and location of the facility as well as references.

The facilities are equipped with diagnostic complexes which are rather different. They contain both the traditional facilities, for example, ordinary impulse ionosonde, and more contemporary analogies (kinesonde, chirpsound) as well as such unique facilities as stations of incoherent scattering. These stations by which the facilities in Arecibo, Tromsø and Khar'kov are equipped are the most informative diagnostic facilities, since they permit the direct study of the plasma and ion acoustic waves (with wave vectors having a projection in the direction of the sounding radio ray), to obtain the height profiles of the density and temperature [*Schlegel et al.*, 1987]. So, considerable plasma displacements (plasma holes) have been studied in modification of the night F-layer of the ionosphere [*Duncan et al.*, 1988]. Again, such effects were also detected in Dushanbe at the beginning of the eighties under definite regimes of modification producing a rather specific picture on the ionograms (of the type of a blossomed-out bed). The chirpsound permits one to obtain a very informative picture of the ionosphere modification (see, for example, *Erukhimov et al.* [1987]). We have mentioned above the method of resonance scattering which permits the observation of a sufficiently large set of ionospheric parameters. We add that this method requires the use of facilities with the impulse power of several hundreds of kW and receiving-transmitting antenna with $G \approx 10^2$. To study the ionospheric turbulence of mean scales, of interest is the method of the signal scintillations from the board transmitters and transit and geostationary spacecrafts. They helped to obtain the basic data discovered in the latter decade on the dynamics of irregularity spectra at different heights from the level of energy release in the modification; in particular, the presence of large vertical velocities (up to 0.5–1 km/s) is shown for the turbulence source propagation from the heating level. Recently, of particular attention regarding the method of the aspect scattering is the investigation of the frequency spectrum scattered by artificial signal irregularities. One can hope that these measurements alone make it possible to determine the character and the values of the motions induced by the modification of both regular and wave (in particular, variations of the electrodynamic drift due to variations of the plasma density in the region of heating) and to find the mechanism of the turbulence spectrum formation (see *Belenov et al.* [1987]). The investigation of peculiarities of the frequency spectrum of combination frequency signals, its polarization and the fine time and spatial structure can be the real method of diagnostics of the natural current system as well as the study of effects associated with their essential modification. In reality, all methods of diagnostics (except for the resonance scattering and CFS) are used, to a certain degree, in the ionospheric investigations. In the case of the ionosphere modification they are applicable; but only as methods of the diagnostics of the modification effects which in their turn can be effective methods for the experimental investigations of the ionosphere.

TABLE I

NN	Name, location geographical latitude	Frequency range MHz	G	PG MW	Year of operation start	Reference
1.	Moscow, 56°	1.35	100	1000 (in pulse)	1961	Shlyuger, 1974
2.	Boulder, 40°	4.5–9 2.7–3.3	60–80 30	120–160 60	1970	Utlaut, 1970
3.	Arecibo, 18°	3–12	200–400	160–320	1971	Gordon et al., 1971
4.	Gorky, 56°	4.6–5.75	100–150	15–22.5	1973	Getmatsev et al., 1973
5.	Monchegorsk, 68°	3.3	130	10	1976	Kapustin et al., 1977
6.	Tromsø, 70°	2.5–8	240	360	1980	Stubbe et al., 1981
7.	Sura(Gorky), 56°	4.5–9	240	150–320	1981	Belov et al., 1981
8.	Gissar (Dushanbe), 38°	3.7–6	60–80	6–8	1981	Erukhimov et al., 1985
9.	Khar'kov, 50°	6–12	150	15	1987	Bogdan et al., 1980

REFERENCES

Belenov, A. F., L. M. Erukhimov, N. A. Mityakov, E. N. Myasnikov and V. L. Frolov, In. Sb. Instabilities and wave phenomenon in ionosphere-thermosphere system, 1987, Gorky, IPF, AN USSR.

Belenov, A. F., V. A. Bubnov and L. M. Erukhimov et al., *Izv. VUZov, Radiofizika*, *20*(12), 1805, 1977.

Belikovich, V. V., E. A. Benediktov, G. G. Getmatsev et al., *Pis'ma v ZhETF*, *22*(10), 497-799, 1975.

Belikovich, V. V., E. A. Benediktov, M. A. Itkina et al., *Izv. VUZov, Radiofizika*, *20*(12), 1821-1826, 1977

Belikovich, V. V., E. A. Benediktov, G. I. Terina, *J. Atmos. Terr. Phys.*, *48*(11-12), 1247-1253, 1986.

Belov, I. F., V. V. Bychkov, G. G. Getmatsev et al., Experimental facility SURA and the first results of investigations on the ionosphere heating, In sb. XXII All-Union Conference on Propagation of Radio Waves, Gorky. Theses of papers. Nauka, M.: 1981, pp. 103-104, June 1981.

Belov, I. F., V. V. Bychkov, G. G. Getmatsev, N. A. Mityakov and G. R. Pashkova, Experimental complex SURA for investigation of artificial ionospheric disturbances, preprint, N167, *NIRFI*, Gorky, 1983.

Belyaev, P. P., D. S. Kotik, S. N. Mityakov, S. V. Polyakov, V. O. Rapoport and V. Yu. Trakhtengerts, *Izv. VUZov, Radiofizika*, *30*(2), 248-267, 1987.

Bespalov, V. I., and V. Yu. Trakhtengerts, Alfven masers, *AN USSR*, Applied Physics Institute, Gorky, 1986.

Bogdan, A. P., A. N. Gridin and A. V. Levkov, Decametric heating facility, in *Investigation of the Ionosphere by the Method of Incoherent Scattering*, no. 170, pp. 44-48, Vestnik, Khar'kov Polytechnical Institute, Kar'kov, 1980.

Duncan, L. M., J. P. Sheerin and R. A. Behnke, *Phys. Rev. Lett.*, *61*(2), 239-334, 1988.

Erukhimov, L. M., V. A. Ivanov, N. A. Nityakov, V. P. Uryadov, V. L. Frolov and V. V. Shumaev, *Izv. VUZov, Radiofizika,*, *30*(9), 1055-1066, 1987.

Erukhimov, L. M., V. Ya. Kovalev, E. P. Kurakin et al., *Izv. VUZOv, Radiofizika*, *28*, 658-662, 1985.

Erukhimov, L. M., S. A. Metelev, E. N. Myasnikov, N. A. Mityakov and V. L. Frolov, *Izv. VUZov, Radiofizika,*, *30*(2), 208-225, 1987.

Gershman, B. N., L. M. Erukhimov, Yu. A. Yashin, *Wave Phenomena in Ionospheric and Cosmic Plasma*, M.: Nauka, 1984.

Getmatsev, G. G., G. P. Komrakov, Yu. S. Korobkov et al., *Pis'ma v ZhETF*, *18*, 621-624, 1973.

Ginzburg, V. L., *Propagation of Electromagnetic Waves in Plasma*, M.: Nauka, 1967.

Ginzburg, V. L., and A. A. Rukhadze, *Waves in Magnetoactive Plasma*, M.: Nauka, 1979.

Grach, S. M., N. A. Mityakov and S. N. Mityakov, Ionosphere disturbance by powerful waves. The results of science and technology. VINITI. ser. *Geomagnetism and High Layers of the Atmosphere*, M.: 1989.

Gordon, W. E., R. L. Showen and N. C. Carlson, *J. Geophys. Res.*, *76*(31), 7808-7813, 1971.

Gurevich, A. V., and A. B. Shvartsburg, *Nonlinear Theory of Radio Wave Propagation in the Ionosphere*, M.: Nauka, 1973.

Kapustin, I. N., R. A. Pertsovskij, A. N. Vasil'ev et al., *Pis'ma v ZhETF*, *25*, 248-251, 1977.

Schlegel, K., M. Rietveld and A. Maul, *Radio Sci.*, *22*(6), 1063-1072, 1987.

Shlyuger, I. S., *Pis'ma v ZhETF*, *19*(5), 247-251, 1974.

Stubbe, P., H. Kopka and R. L. Dowden, *J. Geophys. Res.*, *86*(11), 9073-9078, 1981.

Stubbe, P., et al., *J. Atmos. Terr. Phys.*, *47*(12), 1151-1163, 1985.

Utlaut, W. F., *J. Geophys. Res.*, *75*(31), 6402-6405, 1970.

Yampol'skij, Yu. N., *Izv. VUZov, Radiofizika*, *32*(4), 519, 1989.

Chapter 11

Coherent Radar Techniques

Erling Nielsen
Max Planck Institut für Aeronomie,
D-3411 Katlenburg-Lindau, FRG.

Abstract

At a time when involvement of coherent radars in many research fields is expanding, it is appropriate with a tutorial review of the basic ideas and techniques in use in current coherent radar systems. After introducing the radar principles and some of the properties of radar signals backscattered from the ionosphere, data aquisition, coherent detection, digital sampling and data preccessing are presented. Next the functions of various important elements in a modern radar station are described, and the fundamentals of antenna design with some much used array types are outlined. The capability of coherent radar systems for providing data, that allow estimates of the spatial variations of horizontal electron drift velocities in the ionosphere, is stressed and discussed.

1. Introduction

In most of the plasma regions in the space around the Earth the electrons and ions are to a very good approximation moving under influence of electric and magnetic fields in a so-called ExB drift. This is for example the case in the F-region ionosphere. However, at lower altitudes around 105 km the neutral atmosphere - even though its density is orders of magnitude below that of ions and electrons - has important effects on the mobility of ions, which totally changes the characteristics of the plasma from that at higher altitudes. Because the collision frequency of the ions with the netrual air is larger than their gyro frequency, the ion motion is governed by the electric field and the neutral air. On the other hand, the electron gyro frequency is much larger than the electron collision frequency with the neutral air, and therefore the electron motion is also at these altitudes governed by the electric fields and geomagnetic field, i.e. the electrons atmosphere of the E region that energetic particles from the magnetosphere (at high latitudes) as well as the solar radiation are mainly absorbed, and this results in strong electron density gradients. Thus, in the E region the electrons and ions are moving at a relative speed at an altitude where typically strong density gradients are present. These are the properties that distinguishes the plasma in the 95 to 120 km altitude interval from any other space plasmas associated with the Earth system. The E and F regions contain plasmas of varying properties allowing a varity of fundamental plasma physical processes to occur. Furthermore, from an observational point of view a unique aspect about these space plasmas is, that they are accessible to investigation from the ground using radio waves as well as to in situ investigations using rocket born instruments.

Coherent radars have been and continue to be essential instruments in the study of ionospheric plasma physics. More recently it has been found that coherent radar systems can also be used to determine the horizontal flows of plasma over large areas with good spatial and temporal resolution. Since horizontal plasma flows are governed by the interaction of the solar wind with the Earth's magnetic field and ionosphere, and by other geophysical processes in the magnetosphere and their coupling to the ionosphere, coherent radar systems are now being used to study these processes.

The first observations of coherent radar echoes from the high latitude ionosphere were probably made during the evening hours of December 16, 1938, in Tromsø in northern Scandinavia by Harang and Stoffregen [1940]. Using a pulsed radio wave experiment with a broad beam vertical antenna (a half wave dipole with reflector) they investigated the existence of a hypothetical absorbing layer predicted to be located below the E layer. Operating at 41 MHz at a time of strong magnetic activity they detected weak and variable echoes from an equivalent height of between 400 and 800 km. In hindsight it seems most likely that these echoes originated from magnetic field aligned electron density irregularities in the E- and/or F-layer at ranges, or distances from the antenna, of between 400 and 800 km, i.e. from fairly small elevation angles. During the following five decades electromagnetic waves have been applied extensively to study the properties of the ionosphere using ionosondes and HF, VHF and UHF radars [Fejer and Kelley, 1980]. In particular coherent radars were used to study the equatorial ionosphere [Cohen and Bowles, 1967; Balsley and Farley, 1971; Fejer et al., 1976; Carter et al., 1976; Balsley et al., 1972; Prakash et al., 1974], the high latitude ionosphere [Unwin and Baggaley, 1972; Balsley and Ecklund, 1972; Greenwald and Ecklund, 1975; Tsunoda et al., 1974; Greenwald et al., 1978; Haldoupis and Sofko, 1976; Nielsen et al., 1983a; Greenwald et al., 1985; McNamara et al., 1983], and more recently also the mid-latitude ionosphere [Ecklund et al., 1981;Keys and Andrews, 1984; Riggin et al., 1986].

A radar operating at a certain center frequency, i.e. at a certain wavelength, is sensitive to the corresponding spatial Fourier component of the ionospheric electron density fluctuations along the line of sight of the radar. These spatial fluctuations in the electron density are for some physical processes so long lasting in time that the signals backscattered from a series of pulses have a high coherence. Such processes are detected with a coherent radar. It is of course the physical processes in the ionosphere which cause the signals to be coherent, not the radar, but the basic feature of a coherent radar is such as to take advantage of this coherency . These physical processes are excitation of plasma instabilities which result in plasma waves of relative long durations, typically several milli-seconds. In contrast the incoherent radar relies on backscatter from random, temperature fluctuations of electrons in which case there is no coherency in the signals backscattered from sequential pulses. Incoherent echoes are much weaker than coherent ones, and an incohernt radar is therefore much more complex (and also much more expensive to build) than a coherent radar.

The most fundamental measurement with a coherent radar is the auto-correlation function, or nearly equivalent, the power spectrum of the backscattered radar signals. In order to limit the amount of data output, some radars only measure the intensity of the backscattered signal and (essentially) the first moment of the power spectrum. These parameters can be measured as a function of distance from the radar and of azimuth (at high latitudes) or of elevation (at equator). The width of the antenna beam(s) is typically a few degrees.

Coherent radars have been operated in the interval from 8 to 1210 MHz. The range of peak power used is from 6 to 50 kW. A typical range resolution is 7.5 km, and time resolution varies from a fraction of a second to one and a half minute.

This report is intended to give a general overview and a basic introduction to coherent radar systems to those who consider using such radar systems. In the following sections the basic principles of a pulsed coherent radar and of a continuous wave radar are described, and some observations of ionospheric irregularities are presented. Next the principles of coherent detection, digital measurements and data acuisition are discussed. An example of a new coherent radar system and of different antenna designs are then outlined, and last the principles behind coherent radar measurements of electron drift velocities are discussed.

2. Radar Principles

Coherent radars are operated as pulsed radars or as continuous wave (CW) radars. The name 'coherent' means that the phase of the transmitted signal is preserved in a reference signal with which the received signal is compared. This reference signal is the distinguishing feature of a coherent radar. The principles in these kinds of radars are illustrated in Figure 1.

2.1 Pulse operations

Pulsed radars have been operated with single -, or double - pulses.

2.1.1 Single pulse

Let a pulse of electromagnetic waves of constant peak amplitude and of a duration τ leave the transmitter antenna at time zero, so that time is measured, for example, from the leading edge of the pulse. Ideally the flanks of the pulse should be as steep as possible to make it well defined in time, and therefore its spatial length well defined. However, since many radars are used in populated areas, and therefore should avoid to cause interference, the steepness of the flanks are reduced in order to reduce the bandwidth of the transmitted signal. The rise time at the flank is typically several micro-seconds. A pulse of width τ (measured between the 3 dB points on the flanks) covers a range τ [s] $\cdot$ c [km/s] wide in the direction of wave propagation (where c is the speed of light). If the signal is scattered from space near a distance R_a from the radar, a fraction of the signal reaching that distance will propagate back towards the radar where it arrives at time $t_1 = \frac{2\ R_a}{c}$. Because the backscattered signal from the trailing edge of the pulse, which also arrive at the receiver at time t_1, will have traveled a distance $(R_a - \Delta R) = \frac{(t_1 - \tau)\ c}{2}$, the spatial resolution, range gate or range cell is $\Delta R = \frac{\tau\ c}{2}$.

As the pulse propagates away from the transmitter antenna through a scattering medium, power is continuously backscattered towards the radar, and the final width of the pulse allows a spatial resolution of ΔR, which is obtained when the signal is detected at time intervals equal to the pulse width, τ. This is a typical way of obtaining the variations of backscattered signal power as a function of time, and thus of range. In order to obtain good accuracy of the power it is necessary to average the detected signal over many pulses (forming an ensemble average). It is clear, however, that if several on each other following pulses causes backscatter power to arrive at the radar at the same time, the backscatter

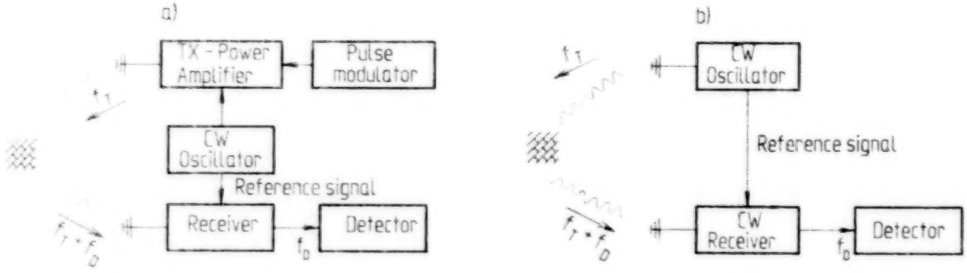


Figure 1: Simple coherent pulse radar (a) and CW radar (b), with emphazis on the feature that makes these radars coherent, the reference signal.

must have occurred at different ranges, and the detected power can therefore no longer be assigned to a certain range. In order to ensure the unambiguious range allocation of a signal detected at a given time, the separation in time between the pulses, the interpulse period, T_{ip}, must be larger than $T_{ip} \geq \frac{2\,(R_{max}-R_{min})}{c}$, where R_{max} (R_{min}) is the maximum(minimum) distance from which backscatter can occur. The inverse of T_{ip} is the pulse repetition frequency, $F_{prf} \leq \frac{c}{2(R_{max}\,-\,R_{min})}$. While R_{max} is dependent only on the extend of the medium through which the pulse propagates, R_{min} is also limited by the pulse width and by how fast the receiver can be activated owing to interference from the transmitted signal (for example owing to ground reflections).

Single pulse operations are also used to obtain power spectra of the backscattered signal as a function of range. Transmitting a series of pulses with equal time spacing, yields for each range gate a time series of backscattered power, the Fourier transform of which is the power spectrum. Since the time it takes to obtain a sufficient long time series for analysis (say 128 ms), is longer than the maximum correlation time of the backscattered signal, this procedure presupposes that the physical process causing the backscatter is constant in time.

If the peak power required for a particular observation program can not be achieved with a practical transmitter for short pulses, then modulated longer pulses can be used together with pulse compression at the receiver, where the received signal is processed in a filter matched to the modulated long pulse. Pulse compression allows a radar to use long pulses to radiate large energy, while simultaneously retaining the good range resolution of the short pulse. This is achieved by use of phase modulation of the radio frequency signal, such that sub-pulses are formed within the long pulse with varying phase of the signal; the phase could for example vary between 0° and 180°. Let a long pulse be subdivided into N sub-pulses of width τ, then the pulse compression ratio is said to be N, and the output of the matched filter will be a spike of width τ with an amplitude N times larger than that of the long pulse. The pulse compression ratio is typically $\gg$1.

2.1.2 Double pulse

Transmitting two pulses with a time distance, T_p, which is an integer times the pulse width ($N\tau$) and (usually) much shorter than the Tip, additional information about the properties of backscattering medium is obtained: the velocity component of the backscattering irregularities along the line of sight of the radar. Assume that backscatter only occurs from a certain range (from a hard target); then even though $T_p < T_{ip}$ the signal received at a given time can be associated with a certain range gate. A coherent radar can in principle detect the Doppler shift in the radar frequency induced by the moving targets, and in practice the radar detects the change in phase of the returned signal from one pulse to the next. This phase change is derived from a comparison between the phase of the returned signal and the phase of the reference signal; the reference signal, in essence, determines the phase of the transmitted signal as a function of time, provided it does not suffer a Doppler shift. On the other hand a phase difference between the reference signal and the received signal then indicates that a moving target has been encountered. If for a moving hard target (for example a copper plate) the phase change is $\Delta\,\phi$ the velocity of the target is $V_d = \text{const.}\ \frac{\Delta\,\phi}{2\,\pi}$. If the plate has moved a quarter wavelength during T_p the distance travelied for the second pulse is a half wavelength longer than for the first, and associated with a phase change of π radians. In that case

$$V_d = V_a = \frac{\lambda}{4\,T_p} \tag{1}$$

Thus, for the general case

$$V_d = V_a \frac{\Delta\,\phi}{\pi} \tag{2}$$

Where V_a is the maximum Doppler velocity, and Pi the largest phase change, one can measure unambiguously with the given values of radar wavelength and pulse separation. The smaller the pulse separation, the larger velocities can be measured.

In reality the backscattering medium is better described as 'soft' and extended in range, rather than 'hard' and limited in range as assumed above. That the target is 'soft' imply that the backscattered signal is no longer discrete in frequencies as for a 'hard' target, but has a spectral width. From the character of the spectrum (or rather the auto-correlation function) it is possible to determine a parameter, which replace the simple form of $\Delta\,\phi$ assumed in the previous paragraph in the calculation of V_d. The broad (in range) backscattering region causes the signal received at a given time to be typically composed of signals from the two pulses backscattered from different ranges. However, by cross-correlating the signal detected at times separated by $N\tau$, it is still possible to obtain the Doppler velocity as a function of range. The signal detection and data handling is outlined in Section 5.

The double pulse procedure can also be used to determine the spectrum of the backscattered signal. This is done by transmitting a series of double pulses, where the pulse separation is increasing from one pulse pair to the next, and the interval between sequential pulse pairs is larger than T_{ip}. One can show that each pulse pair determines one point on the auto-correlation function. The Fourier transform of the auto-correlation function is the power spectrum. The number of double pulses used has typically been 10, and for T_{ip} = 10 ms , it takes 100 ms to obtain one complete auto-correlation function for analysis. Since this time is longer than the maximum correlation time of the backscattered signal, this procedure also presupposes that the physical process causing the backscatter is constant in time.

2.1.3 Interferometer

The backscattered radar signal does not only contain information about the backscatter cross section, Doppler velocity and auto-correlation function of the backscattering medium, but also more detailed information about the direction to the backscattering region [Ierkic, 1980; Farley et al., 1981; Providakes et al., 1983; Providakes, 1985]. Assuming the backscatter originates in discreet backscatter centers, the distances travelled from such a center to two antennas, horizontally separated by typically 10s of meters, will differ by a fraction of a wavelength (plus a whole number of wavelengths). Thus, the phase difference of the signal received at the two antennas indicates the direction to the backscatter centers. If the connecting line between the antennas is perpendicular to the direction of the antenna lobes, then the azimuthal angle from the normal to the line directed towards the backscatter irregularity can be measured. If the connecting line between the antennas is parallel to the antenna lobes, then the elevation angle to (essentially the height of) the backscatter centers can be measured.

2.2 Continuous wave (CW)

A CW system have been used in which the transmitter and receiver arrangements are separated by, say, 40 km along the east-west direction. East-west directed antennas are used to transmit and receive the reference signal needed for the data handling. In this configuration the system is sensitive to north-south motions of the targets. One outstanding advantage of this technique is the very high time resolution, typically 1 ms, with which one can obtain power spectra of the backscattered signal.

3. Ionospheric irregularities

The volumen cross-section of backscatter from the ionosphere is determined by the mean square electron density fluctuations, $< \Delta N^2 >$, and the (normalized) spatial power spectrum, f(k), of the density fluctuations [Farley et al., 1981] ,

$$\begin{aligned} \sigma(\underline{k}_w) &= 4\pi r_e^2 \int < \Delta N(\underline{r}) \Delta N(\underline{r}+\underline{r}') \, e^{i\underline{k}\cdot\underline{r}'} \, d^3\, \underline{r}' \\ &= 32\pi^4\, r_e^2 \, < \Delta N^2 > f(\underline{k}_w) \end{aligned} \tag{3}$$

where $\underline{k}_w$ is the wave vector, and r_e is the classical electron radius.

From the momentum equation [Fejer and Kelley, 1980],

$$\underline{k}_t = \underline{k}_w + \underline{k}_s \tag{4}$$

where t, w, and s stands for transmitted-, wave- and scattered signals. In the case of backscatter $\underline{k}_s = -\underline{k}_t$, and therefore, $k_w = 2\ k_t$. Thus, the wavelength of the electron density fluctuations (the plasma waves) probed by the radar is equal to half the radar wavelength.

Electron density fluctuations in the E region arise owing to exitation of plasma instabilities. When the electron drift velocity, V_e = ExB, in the ionosphere exceeds the ion acustic velocity and/or when the scale length, L_e, of electron density gradients is not much larger than the signal wavelength, then plasma waves can be excited and the ionospheric plasma break up into small scale electron density structures. The wave frequency, ω_w, for plasma waves propagating nearly perpendicular, i.e. within a few degrees, to the geomagnetic field, at altitudes where the ions may be considered unmagnetized, for electrons nearly free of collisions with other particles, and for wave growth rates, γ, small compared to ω_w, has the form [Fejer and Kelly, 1980; Fejer et al., 1984],

$$
\begin{aligned}
\omega_w &= \frac{\underline{k}_w \cdot (\underline{V}_e + \psi \underline{V}_i)}{1+\psi} \\
\gamma &= \frac{1}{1+\psi}\left[\frac{\psi}{\nu_i}\left(\omega_r^2 - k_w^2 C_s^2\right) + \frac{\omega_r \nu_i}{k_w L_e \Omega_i}\right] - 2\alpha N_o \\
\psi &= \psi_o\left(+ \frac{\Omega_e^2}{\nu_e^2}\frac{k_{w,11}^2}{k_w^2}\right) \\
\psi_o &= \frac{\nu_e \nu_i}{\Omega_e \Omega_i}
\end{aligned}
\tag{5}
$$

and the velocity at threshold ($\gamma = 0$) is

$$
\begin{aligned}
\underline{k}_w \cdot \underline{V}_d &= k_w C_s (1+\psi) F^2 + 1 + \frac{2\alpha N_o \nu_i (1+\psi)}{k^2 C_s^2 \psi}^{1/2} - F \\
F &= \frac{\nu_i^2}{2\Omega_i k_w^2 C_s L_e \psi}
\end{aligned}
\tag{6}
$$

where $\nu_e(\nu_i)$ and $\Omega_e(\Omega_i)$ are the electron(ion)-neutral collision frequencies and the electron(ion) gyro frequencies, respectively. $V_e(V_i)$ are the electron(ion) drift velocities, and C_s the ion-acoustic velocity. The last term in the expression for the growth rate accounts for damping owing to recombination. α is the recombination coefficient and N_o is the mean electron density. For typical values of the parameters at the altitude of the backscattering layers ψ is much smaller than one.

In the absence of density gradients, $L_e = \infty$, (short wavelength approximation) the growth rate becomes positive when the wave phase velocity $\frac{\omega_w}{k_w}$ is larger than C_s, and so much larger as to also overcome the recombinational damping. This requirement will first be satisfied for small values of the angle Arccos $\left(\frac{\underline{k}_w \cdot \underline{V}_e}{k_w V_e}\right)$, i.e. inside a cone centered on the direction of the electron drift velocity. The approximate opening angle of the cone is Arccos $\left(\frac{C_s}{k_w V_e}\right)$. Waves propagating in this cone are called Farley-Buneman waves. If these (primary) Farley-Buneman waves are propagating across an electron density gradient, which has a component parallel to the DC electric field, then polarization electric fields, E_p, are formed inside the primary wave structure in directions parallel to the DC electron drift velocity. The polarization field together with the density gradients associated with the primary waves may cause the second term in the expression for the growth rate to be large enough to make γ positive, that is to excite secondary waves. The value of V_e to be used in Equation 5 is E_p x B, and thus, the phase velocities of the secondary waves will be directed at large angles to the direction of the DC eletron velocity, essentially filling the angular interval outside the Farley-Buneman cone.

On the other hand, even if the DC electron drift velocity is smaller than the ion-acoustic velocity the second term of the growth rate may still cause γ to be positive. This can occur if k_w L_e is sufficient small, i.e. if Le is not too large compared to the wavelength (long wavelength approximation). In this case γ may no longer be small compared to the wave frequency, which then has the form [Kudeki et al., 1982],

$$\begin{aligned} \omega'_w &= \omega_w \frac{1}{1+\left(\frac{k_o}{k_w}^2\right)} \\ k_o &= \frac{\nu_i}{\Omega_i}\frac{1}{1+\psi}\frac{1}{L_e} \end{aligned} \tag{7}$$

This mechanism is known to operate in the equatorial ionosphere to excite long wavelength waves. Furthermore, these (primary) waves may have associated polarization electric fields (directed horizontally) large enough to excite (secondary) Farley-Bunemann waves (propagating vertically).

Waves can also be excited in plasmas at altitudes where the ions are magnetized, $\Omega_i \gg \nu_i$, [for example Fejer et al.,1984], in which case the wave frequency at marginal instability is

$$\omega_w^2 = \Omega_i^2 + k_w^2 C_s^2 \tag{8}$$

The ion cyclotron instability is generated by electrons flowing parallel to the geomagnetic field (parallel currents). The waves propagate on a cone centered on the magnetic field, and the phase velocity has a component both parallel to and perpendicular to the magnetic field. In regions, like the upper part of the auroral electrojet, where $\Omega_i = \nu_i$ the Farley-Bunemann - and the ion cyclotron instability may be simultaneously excited.

Figure 2 displays some backscatter spectra observed at high latitudes at 140 MHz. In panel a the typical spectra are shown. The thin curve is the power spectrum of the primary Farley-Bunemann waves, and the thick curve is the spectrum of the associated secondary waves. Panels b and c show unusual spectra with very small (b) and very large (c) velocities at the spectral peaks [Haldoupis and Nielsen, 1989]. Figure 3 shows some of the observed characteristics of the signal-to-noise ratio and the mean Doppler shift of the backscatter signal as a function of aspect angle, flow angle and electron drift velocity. The aspect angle is the angle between $\mathbf{k}_w$ and the direction of the geomagnetic field. The flow angle is the angle between $\mathbf{k}_w$ and the electron drift velocity. Panel a is in the plane perpendicular to the magnetic field and displays the variations of the mean Doppler velocity as a function of the flow angle [Reinleitner and Nielsen, 1985]. At large angles it follows a cosine law, that is the phase velocity in the direction of $\mathbf{k}_w$ equals the component of the electron drift on that direction. At smaller angles, inside the Farley-Buneman cone, the phase velocities are limited in magnitude, at values, as shown in panel b, near the ion acoustic velocity (limited by the concentric circles). The strong dependence of the phase velocities of the primary waves on the aspect angle is demonstrated in panel c [Nielsen, 1986]. In the next two panels the strong variations of the signal-to-noise ratio for both primary and secondary waves are shown as a function of the flow angle (d) and aspect angle (e) [Andre,1983; Nielsen,1988]. The relative mean square electron density fluctuations are seen, in panel f, to increase with the electron flow velocity, and to level off, probably to saturate, at larger speeds [Nielsen et al., 1988].

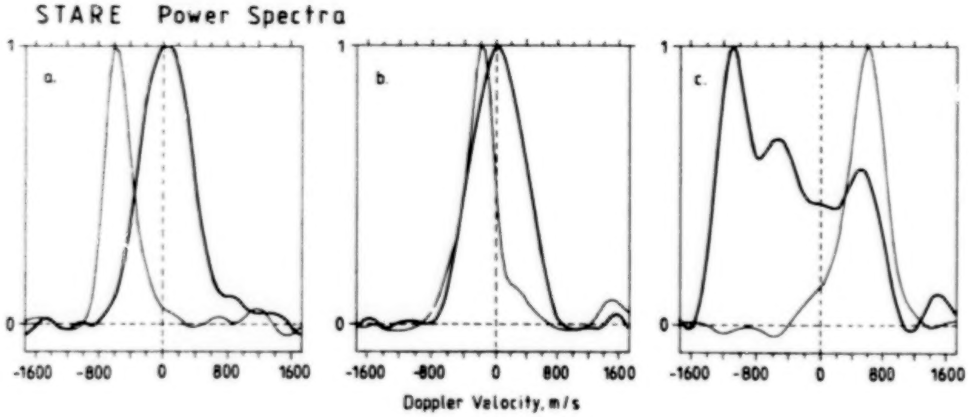


Figure 2: Backscatter spectra in the auroral electrojet. Panel a shows typical examples for observations at small - (thin line) and large (thick line) flow angles. In Panel b both mean Doppler velocities are small, below the ion acoustic velocity, while in Panel c the Doppler velocity in one spectral peak is much larger than any realistic ion acoustic velocity [from Haldoupis and Nielsen, 1989].

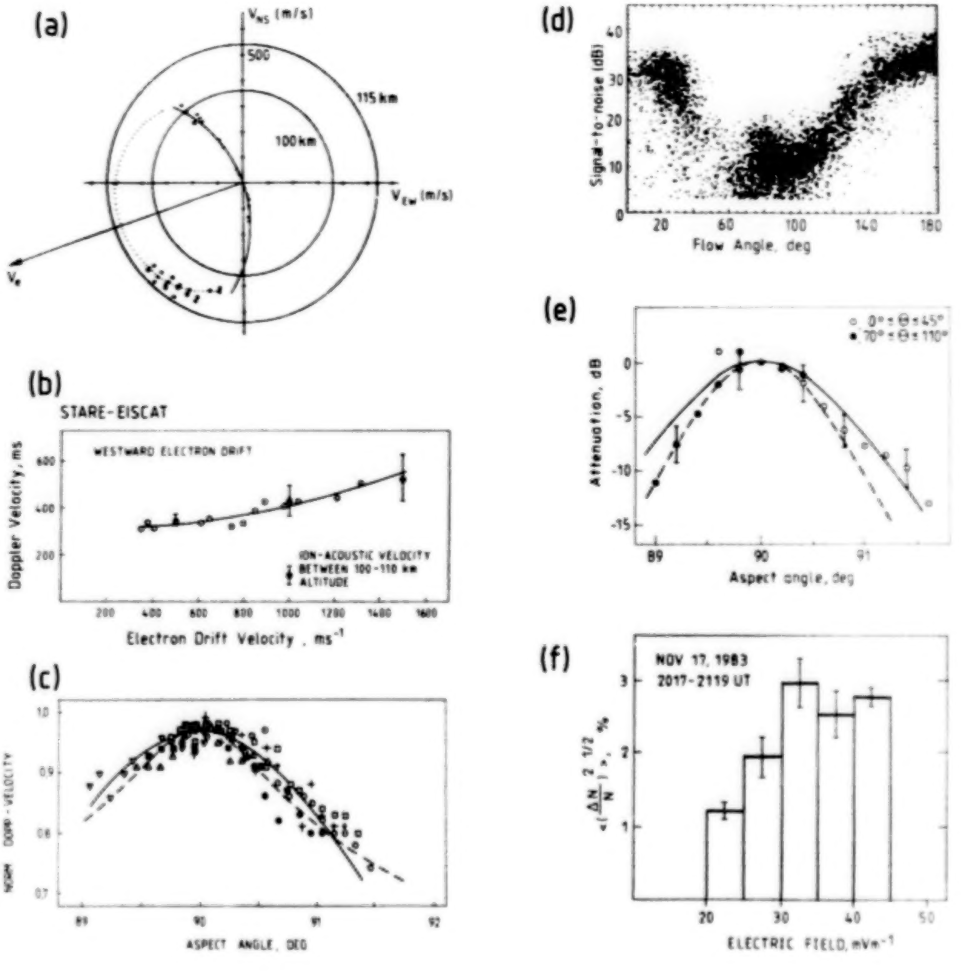


Figure 3: Some characteristic properties of the mean Doppler velocity and the signal-to-noise ratio as functions of electron drift speed, flow angle and aspect angle (see text).

4. The Radar equation

If the power of a radar transmitter is denoted P_t, and the gain of the transmitting antenna is G_t, and its radiation pattern is $g_t(\theta, \phi)$, where θ and ϕ are the horizontal and elevation angle, then the power density of the transmitted signal at a distance R in direction (θ, ϕ) is

$$R_R = P_t \frac{G_t \; g_t(\theta, \phi)}{4\pi \; R^2} \tag{9}$$

Let the radar backscatter cross section per unit volume be σ, the backscatter volumn V_s, and the effective receiving area of the receiver antenna A_e, then the received power, P_r, at the antenna is

$$P_r = P_t \frac{G_t g_t(\theta, \phi)}{4\pi R^2} \frac{\sigma V_s}{4\pi \; R^2} A_e \tag{10}$$

where the effective area, A_e, is defined by

$$A_e = \frac{G_r \lambda^2}{4\pi} \tag{11}$$

here G_r is the gain of the receiver antenna.

The antenna beam at the range of the backscatter region is normally not filled out with backscatter centers. At high latitudes V_s extends across the beam in the horizontal direction while it typically may have vertical thickness, H, of 10 km. If the antenna beam width is $\Delta\theta$, then V_s can be written

$$V_s = \frac{R\Delta\theta \;\; H_c \tau}{2} \tag{12}$$

where $\frac{c\tau}{2}$ is the depth of the backscatter volumn. The received power can now be written

$$P_r = P_t \frac{G_t \; g_t}{4\pi \; R^2} \cdot \frac{\sigma}{4\pi \; R^2} \cdot \frac{G_r g_r \lambda^2}{4\pi} \cdot \frac{R\Delta\theta \; H_c \tau}{2} \tag{13}$$

where g_r is the (normalized) angular radiation pattern of the receiving antenna. The radar scattering cross section, σ, is a function of frequency or k-vector. (See also Walker et al.[1987] for a detailed treatment of the radar equation).

Moorcroft (1987) used observations from several radars operating in the range from 50 to 800 MHz to determine the relationship $\sigma(k)$ for the strongest observed signals,

$$\sigma(k) = 3 \cdot 10^{-7} \cdot k^{-2.25} [m^{-1}] \tag{14}$$

This result is shown as a solid line in Figure 4. The radar cross sections at 30 MHz were considerably larger, and at 1210 MHz lower than predicted by extrapolating from the above equation.

The figure also indicates that the possible cross section values, for a given frequency, range over 4 to 5 order of magnitudes. This spread in radar cross section results not only from its the dependence on many ionospheric parameters, but also from its depependence on the geometry between the directions of the radar line of sight, the geomagnetic field and the ionospheric electron drift velocity, as outlined in the previous section.

5. Data handling

5.1 Coherent detection

The transmitted signal of frequency ω, phase ϕ, and amplitude a(t), where the time variations refere to pulsed operations, has the form

$$s(t) = a(t) exp(\omega t + \phi(t)) \tag{15}$$

The signal arriving at the receiver antenna is a result of backscatter processes in the ionosphere and noise, can be written

$$z(t) = a_1(t) cos(\omega t) + i \quad a_2(t) sin(\omega t) \tag{16}$$

This signal is amplified in the receiver and then coherently detected, as shown in the simplified schematic drawing in Figure 5.

By mixing the received signal with the reference signal, exp (ωt), phase shifted 0° and 90°, two output channels, the in-phase and the quadrature output, respectively, are formed

$$\begin{aligned} x(t) &= A(t) cos[\phi(t)] \\ y(t) &= A(t) sin[\phi(t)] \end{aligned} \tag{17}$$

where

$$\begin{aligned} A(t) &= \frac{1}{2}\sqrt{a_1^2(t) + a_2^2(t)} \\ \phi(t) &= Arctan\left(\frac{a_2(t)}{a_1(t)}\right) \end{aligned} \tag{18}$$

The bandwidths of the transmitted signal and of the receiver are both larger than the spectral width of the received signal. The results of deconvoluting the observed signal with these instrumental effects can therefore normally be disregarded.

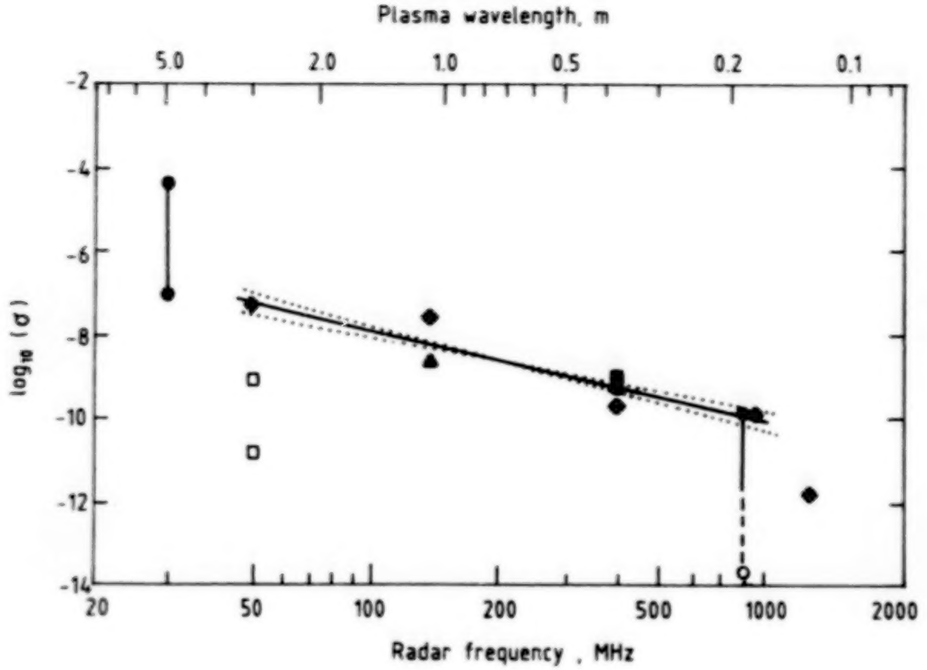


Figure 4: Experimentally determined variations of the maximum volume backscatter cross section in the auroral region (black signature) as function of radar frequency and plasma wave length. The white signature is used for equatorial observations [from Moorcroft, 1987].

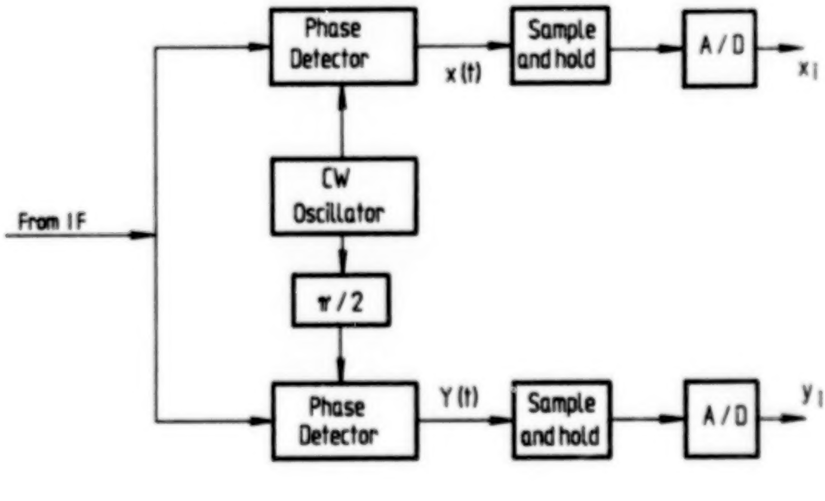


Figure 5: Simple block diagram of signal processing in coherent detection, producing the in-phase and the quadrature output, xi and yi, respectively.

5.2 Digital measurements

The time variations of the received signal is obtained by digitally sampling the input with constant time intervals equal to the pulse width. The i'th sample has the form

$$\begin{aligned} x_i &= A_i cos\phi_i \\ y_i &= A_i sin\phi_i \end{aligned} \tag{19}$$

where $A_i(\phi_i)$ is the amplitude (phase) of the received signal. Conventionally the received signal and its phase are written,

$$\begin{aligned} z_i &= x_i + iy_i \\ \phi_i &= Arctan\left(\frac{y_i}{x_i}\right) \end{aligned} \tag{20}$$

Independent of the transmitted pattern of pulses all information about the backscattering medium is contained in the so-called 'correlation matrix' [Turunen and Silen, 1984]. The matrix has the dimension $M \cdot M$, where M is the maximum number of samples, and the matrix elements have the form $c_{ij} = z_i z_j^*$, where i=1,M and j=i,M. In the general case not all the matrix elements are of interest, many will equal zero because the signal at time 'i' and 'j' are uncorrelated. Examples of correlation matrices for single pulse, double pulse and multi(4) pulse operations are shown in Figure 6a,b and c.

5.3 Single pulse

For a single pulse the power as a function of time or range (of i) has the form,

$$P_i = c_{ii} = x_i^2 + y_i^2 \tag{21}$$

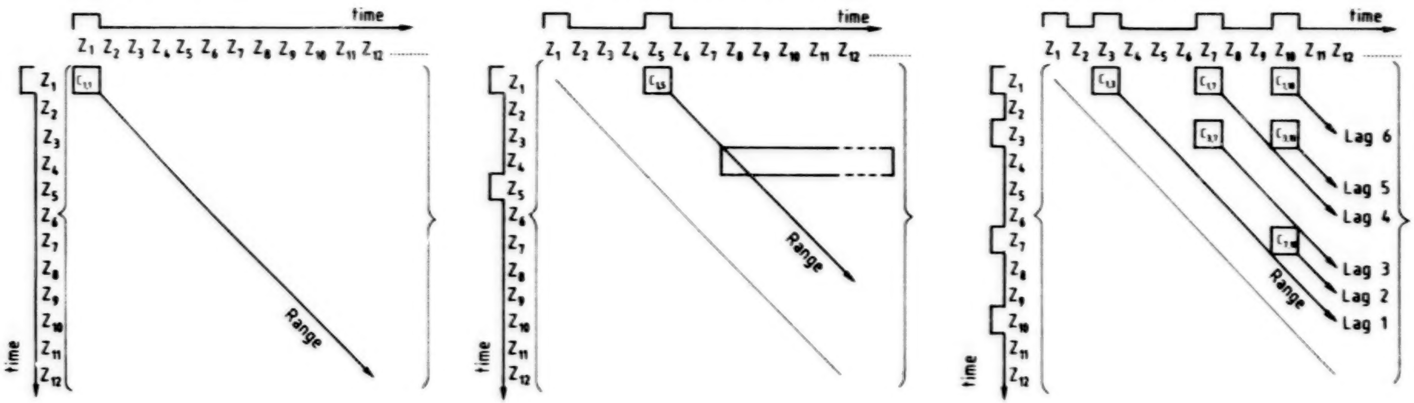
In the correlation matrix the power as a function of range is located on the diagonal (Figure 6a). During one integration time (typically 20 seconds) a number of single pulses, n_{ave}, are transmitted; the average power from the i'th range is

$$< P_i > = \frac{1}{n_{ave}} \sum_{k=1}^{n_{ave}} (P_{ik} - Noise_k) \tag{22}$$

The Noise sample are measured after the pulse has passed the maximum range from which backscatter can occur.

5.4 Double pulse

Now, following Rummler [1968], assume that two pulses separated by an integer, N, times the pulse width are transmitted (for a pulse width of 100 micros the pulse separation may be 300 micros, i.e. N=3); the received signal at times i and j=i+N are



a) Single Pulse (Power)

b) Double Pulse (Doppler Velocity)

c) Multi Pulse (ACF)

Figure 6: The correlation matrix is illustrated for three different types of pulse patterns: single-, double-, and multi pulse. Each element in the matrix is the product of two samples taken at the same or at different times. Only those elements are different from zero for which a part of the sampled signal originated at the same range. Thus for a single pulse only $c_{ii} \mp 0$; for a double pulse with pulse separation 4τ only $c_{i,i+4} \mp 0$; for the case of a multi pulse the situation becomes more complicated as shown in Panel c. The range variation for each lag follows lines parallel to the diagonal. It is clear that only such pulse patterns are allowed for which all the 'lag-range'- lines do not overlap.

$$\begin{aligned} z_i &= x_i + iy_i \\ z_j &= x_j + iy_j \end{aligned} \tag{23}$$

and the phase difference at these two times is

$$\phi_i - \phi_j = Arctan\left(\frac{y_i}{x_i}\right) - Arctan\left(\frac{y_j}{x_j}\right) \tag{24}$$

Now consider the multiplication of the two signals z_i and z_j^*. One obtains,

$$\begin{aligned} c_{ij} = z_i z_j^* &= (x_i + iy_i)(x_j + iy_j) \\ &= (x_i\, x_j + y_i\, y_j) + i(x_j\, y_i - x_i\, y_j) \end{aligned} \tag{25}$$

and noting that

$$\begin{aligned} Re(z_i\, z_j^*) &= x_i\, x_j + y_i\, y_j \\ &= A^2(cos\phi_i cos\phi_j + sin\phi_i sin\phi_j) \\ &= A^2 cos(\phi_i - \phi_j) \\ Im(z_i\, z_j^*) &= x_j\, y_i - x_i\, y_j \\ &= A^2(cos\phi_j sin\phi_i - cos\phi_i sin\phi_j) \\ &= A^2 sin(\phi_i - \phi_j) \end{aligned} \tag{26}$$

the phase difference can be written as

$$\phi_i - \phi_j = Arctan\left(\frac{Im(z_i\, z_j^*)}{Re(z_i\, z_j^*)}\right) \tag{27}$$

and the angle of the product of the signals received at times i and j is

$$\begin{aligned} ang(z_i\, z_j^*) &= Arctan\left(\frac{Im(z_i\, z_j^*)}{Re(z_i\, z_j^*)}\right) \\ &= \phi_i - \phi_j = \Delta\phi \end{aligned} \tag{28}$$

Inserting this value of the phase change in Equation 2 the target velocity results,

$$V_d = V_a \, \frac{ang(z_i\, z_j^*)}{\pi} \tag{29}$$

Since $c_{ij} = c_{i,i+N}$ these elements, or Doppler velocities, are on a line parallel to the diagonal in the correlation matrix, with the velocity corresponding to the first range located in the first row and fourth column (Figure 6b).

Now assume a broad backscatter region, i.e. that at a given time the received signal is composed of backscatter from different ranges, as illustrated in Figure 7.

The signals at time t_i and t_j are

$$\begin{aligned} z_i &= z(t_i) = S_{ia} + S_{ib} = x_{ia} + iy_{ia} + x_{ib} + iy_{ib} \\ z_j &= z(t_j) = z(t_i + T_p) = S_{jc} + S_{ja} = x_{jc} + iy_{jc} + x_{ja} + iy_{ja} \end{aligned} \tag{30}$$

where indices a,b and c indicate the ranges, R_a, R_b and R_c, from which backscatter originates. The product of the signal at these two times is

$$\begin{aligned} z_i\, z_j^* &= z(y_i)z^*(t_i + T_p) \\ &= S_{ia}S_{jc}^* + S_{ia}S_{ja}^* + S_{ib}S_{jc}^* + S_{ib}S_{ja}^* \end{aligned} \tag{31}$$

Because the two fractions of a signal which originates at different ranges (say R_a and R_b) are uncorrelated (because they are spatially so far apart; for $T_p = 300$ micros, 45 km), it follows that the ensemble average, the time average over the integration time, of this product is equal to the average of those terms originating at the same range, so

$$\begin{aligned} < z_i\, z_j^* > &= < z(t_i)z^*(t_i + T_p) > \\ &= < (S_{ia}\, S_{ja}^*) > = < (x_{ia} + iy_{ia})(x_{ja} - iy_{ja}) > \\ &= \frac{1}{n_{ave}} \sum_{k=1}^{n_{ave}} z_{ik}\, z_{jk}^* \end{aligned} \tag{32}$$

In the time averaging of $z_i\, z_j^*$ it is the averages of the real and imaginary components of the product which are formed. In the general case the backscatter intensity from ranges R_b and R_c are as strong as from range R_a. Since the signal from the two former ranges are unwanted, i.e. enter the calculations as noise, 50 % of the signal measured at a given time is noise, and so the signal-to-noise ratio is no better than 3 dB.

With Equation 32 the target velocity at range R_a, also for a soft target, can be written as Equation 29, and since the value of the auto-correlation function of the received signal at time T has the form $R(T) = < z(t)z(t+T) >$, V_d can be expressed as

$$V_d = V_a \frac{ang[R(T_p)]}{\pi} \tag{33}$$

A given target velocity, v, will cause a given Doppler shift, f, of the radar frequency, which is related to the phase change, $\Delta\phi$, in the interpulse time, Δt, as

$$2\pi f = \frac{\Delta\phi}{\Delta t} = \frac{2\pi\Delta r}{\lambda\Delta t} = \frac{4\pi v}{\lambda} \tag{34}$$

$$f = \frac{2v}{\lambda} \tag{35}$$

At for example 140 MHz the wavelength is $\lambda = 2.14$ m, and thus f = 1.07 v.

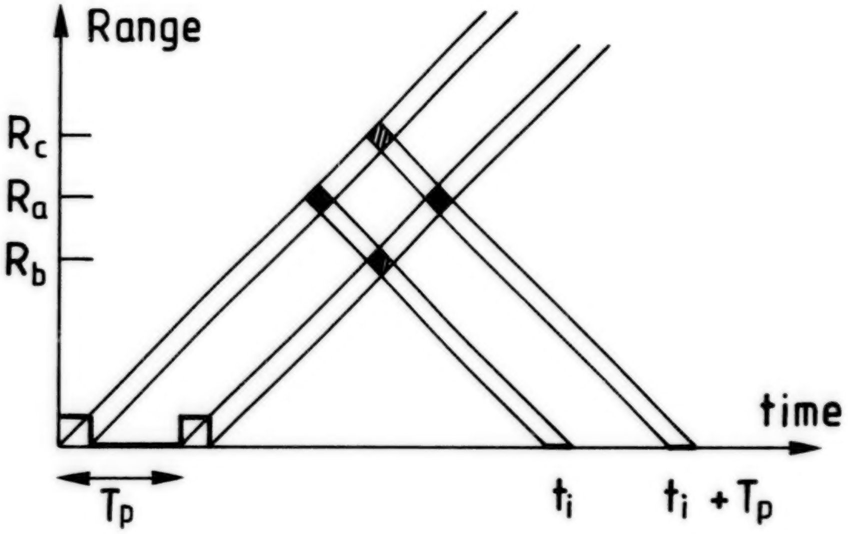


Figure 7: Range-Time diagram for a double pulse. It is illustrated that the signals received at t_i and $t_i + T_p$ both contain backscatter signal from range R_a. The two signals z_i and $z_i + T_p$ will therefore have a correlating part, and the time average of the product $z_i z_i^* + T_p$ will be representative of the backscatter at range R_a.

V_d calculated according to Equation 33 is related to the mean Doppler frequency, $< f >$, of the backscattered power spectrum S(f). $< f >$ is defined by the expression

$$\int S(f)(f - < f >)df = 0 \tag{36}$$

where S(f) is normalised so that

$$\int S(f)df = 1 \tag{37}$$

The relationship between the auto-correlation and the spectrum is

$$R(T) = \int S(f)\exp(i2\pi fT)df \tag{38}$$

and introducing the definition $ang(R(T)) = 2\pi f_o T$, use $R(T)$ can be written

$$R(T) = |R(T)|\exp(i2\pi f_o T) \tag{39}$$

and therefore

$$Im[R(T) \cdot \exp -(i2\pi f_o T)] = 0 \tag{40}$$

thus, combining Equations 38 and 40 yields,

$$\int S(f)sin(2\pi T(f - f_o))df = 0 \tag{41}$$

This equation is satisfied for symmetric spectra, as well as for spectra which are only significantly larger than zero where the argument to the sine-function is small, in which case

$$\int S(f)(f - f_o)df = 0 \tag{42}$$

Comparing Equations 41 and 42 with Equation 36 indicates that $f_0 =< f >$ in these cases, i.e. the double pulse technique yields a good estimate of $< f >$ for symmetric spectra, or for any spectral shape if the requirement T (f-f_0) $\ll$1 is satisfied. This requirement imply that the spectrum must be narrow compared to $1/T_p$ (for example: f-$f_0 \ll 1667$ Hz for T=300 micros). In practice we find, at 140 MHz, that the 'mean' Doppler velocity obtained using the double pulse technique is typically about 15 % smaller than the actual mean frequency of the power spectrum.

It was noted above that the signal-to-noise ratio for Doppler velocity measurements typically was 3 dB, independent of the gain in the radar system. The question naturally arise, whether or not a different technique to performe these measurements could be devised such as to improve the statistics of the result for a given integration time, or to shorten the integration time and still obtain a certain accuracy. A possible way would be to mark the two pulses using different frequencies. Assuming the transmitter and antennas had bandwidths wide enough to pose no limitations, the receiver could filter the signal to separate it into components associated with the respective frequencies. These components would then be coherently detected in separate detectors, yielding for each sample two complex outputs, one associated with the first pulse only, z_{i1}, and one with the second pulse only, z_{i2}. Using z_{i1} and z_{j1} in calculating V_d would considerably increase the signal-to-noise ratio because contributions to the signal from unwanted ranges (R_b and R_c) would be eliminated leading to quicker convergence of the time average, allowing shorter integration times to be used.

The recognition that the time average of the received signals at time t and t + T_p yields the value of the auto-correlation function for $T_p = N\tau$, the double pulse separation, it is clear that the complete auto-correlation function can be determined by using a series of double pulses with increasing pulse separation (the spectra in Figure 2 were obtained in this way). In the framework of the correlation matrix it is easy to see that the elements of auto-correlation function are located in the horizontal rectangle (in Figure 6b arbitrarily shown for the fourth range). The time required using the Stare radar system to obtain one estimate of 11 points (lags) on the auto-correlation function is, measuring for ranges up to 1230 km, about 300 ms (Nielsen et al.,1984).

5.4 Multi pulse

The rather detailed treatment of the douple pulse case has shown that the time average of the signals detected at time intervals equal to the pulse separation yields the important information about the signal, that is points on the auto-correlation function. In the very general case one can see that all information about the backscattering process, i.e. the auto-correlation function, as a function of range is contained in the products one can form of the time series of the digitally sampled signals. For a 4 pulse code ($n_p = 4$) one can measure the auto-correlation function at $n_l = 6$ lags ($n_l = n_p(n_p - 1)$). The correlation matrix for a 4 pulse code is illustrated in Figure 6c. The squares mark the locations of the matrix elements, which belong to the first range, and the locations of the elements with range are marked by the lines parallel to the diagonal. Farley [1972] has considered a number of multi-pulse patterns that might be used for radar auto-correlation function studies. Multi-pulses are currently being used in HF radar studies of the polar ionosphere [Greenwald et al., 1984].

It would seem concivable that the n_e-lag auto-correlation function could be measured faster and with the same accuracy if instead of double pulses, several pulses with interpulse separations (separations between two arbitrary pulses) from τ to Nτ were transmitted together. It is also clear why this is not necessarily the case: because several pulses are transmitted the backscattered signal arriving at the radar at a given time is composed of parts from, in the general case, each of the pulses, and not only from the one pulse located at the wanted range. Thus, if for example seven pulses are transmitted into a homogeneous backscatter region the received signal has a signal-to-noise ratio of 7/6, or 0.7 db; much reduced from the double pulse case. To obtain a certain accuracy more estimates of the auto-correlation function must therefore be made in the multi pulse case than for the double pulse case. Farley [1969] determined the consequences of this, and found the integration time to be proportional to

$$R_t = \frac{(n_p + N/S)^2}{n_p(n_p - 1)} \tag{43}$$

With the noise-to-signal ratio, N/S = 0, one finds for n_p = 2 (1 lag) R_t = 2, and for n_p = 5 (10 lags) R_t = 1.25. Thus, the multi pulse (5) integration time (for a given accuracy estimating the correlation) is 0.63 times the double pulse integration time. Instead of an integration time of 30 s using double pulses, for the case mentioned in the end of the previous section, one can obtained the correlation with the same accuracy in 19 s using a 5-pulse pattern.

It is speculated that the use of an 'alternating' pulse pattern [Lehtinen and Haggstrom, 1987] can even further reduce the integration time [Lehtinen, private communication].

5.5 Interferometer

In an altitude interferometer there are two antennas located at heights h_1 and h_2, separated by distance d and lined up in the direction of the line-of-sight. Owing to this spatial separation of the antennas the phase difference, $\Delta\phi$, between the two antennas, expressed in terms of the elevation angle, 90-θ, to the backscatter centers, is [Providakes, 1985]

$$\Delta\phi_{12} = kdcos(90 - \theta) + Arctan\left(\frac{sin2a_1 + 2sina_1cosa_2}{1 + cos2a_1 + 2cosa_1cosa_2}\right) \tag{44}$$

where it is assumed that the ground is an ideal reflector (a good approximation at 140 MHz), and where

$$\begin{aligned} a_1 &= k(h_1 - h_2)sin(90 - \theta) \\ a_2 &= k(h_1 + h_2)sin(90 - \theta) \end{aligned} \tag{45}$$

For a double pulse with time separation Nτ one point on the complex cross-correlation function of the signals, $z_{i,1}$ and $z^*_{i+N,2}$ (where the subscripts 1 and 2 refer to the antennas), is obtained by forming the time average of the product $z_{i,1}\, z^*_{i+N,2}$. Using a series of double pulses with increasing pulse separation, or a multi pulse pattern, determines the complex cross correlation function of the backscattered signal received at the antennas. The phase of the cross correlation function, i.e. the phase difference at the antennas for the i'th range , is then given by

$$\Delta\phi_{i,12} = Arctan\left(\frac{Im(z_{i,1}z^*_{i+N,2})}{R_e(z_{i,1}z^*_{i+N,2})}\right) \tag{46}$$

Equations 44 and 46 determine the zenith angle, and thus the altitude of the backscatter centers (within the uncertaincy that the phase only is measured between 0° and 360°).

6. A Coherent Radar System

Stare (Scandinavian Twin Auroral Radar Experiment) is a coherent radar system consisting of two identical radar stations, which have been in operation since 1977 [Greenwald et al., 1978]. It is currently in the process of being up-graded, and has therefore been chosen here as an example of a modern radar system. The system can be used as an instrument to study plasma physics of the ionospheric E region as well as to measure electron drift velocities over a large area in the ionosphere.

6.1 A Radar station

Figure 8 is a block diagram of a Stare radar station. The station consists of a 50 kW transmitter which can transmit pulses 100 or 50 micros wide. These the RF signal may also be switched in phase from 0° to 180°. The signals are radiated from a broad beam antenna at low elevation angle towards the auroral zone. The backscattered signals are pre-amplified and detected in 9 independent directions, in 9 narrow antenna beams, then amplified, filtered and coherently detected in the receiver unit. The analog quadrature output from the nine channels are digitised in the correlator, a unit with the main purpose of forming the correlation matrix, that is forming the ensemble averages (Equation 32) of the (relevant) matrix elements for each channel. The output of the correlator, the auto-correlations (or maybe only the power) as a function of range for the nine channels, are written to the data storage devices, optical disks, together with a header, containing information about the time of the measurements, pulse patterns used, and the field of view, as well as housekeeping data indicating the state of the pre-amplifiers, the transmitted power and the reflected power (owing to a possible mismatch between antenna and transmitter).

The process-computer allows an operator to define the radar operations, and it then controlls these operations. At a console an operator can input the values of the physical quantities, which defines the operations (for example: near - and far range as well as the azimuthal coverage of the field of view, the pulse patterns (say a single pulse followed by multi pulse, noise measurements, and integration time). A software package translates these quantities into radar operating parameters and writes them into two files . These files, essentially, establishes the time sequence of the operations the radar has to perform, and the operations the correlator has to perform including the selection of the correlation matrix elements relevant for the chosen pulse patterns. The 'time sequence' file is copied to the memory of the radar controller and the other file to the correlator. A detailed technical description of the controller and correlator is in Schmidt et al. [1988].

The settings of the transmitter and receivers, and the time sequence of radar operations are controlled by the radar controller. It has 9 outputs going to the transmitter, receiver and the correlator. These outputs controll and/or set parameters before the start of measurements, before onset of an integration time, or during a scan (i.e. during the transmission of a group of pulses and the receiving of the backscattered signals). Before an experiment the transmitter power level is set, based on the given pulse pattern and scan time, to ensure the duty cycle of the transmitter is not exceeded, and the receiver filters are set to match the pulse width. At the end of an integration time a message is send to the correlator to start transfering the measured auto-correlation functions via the process computer to the optical disks, and the gain of the 9 receiver channels are set

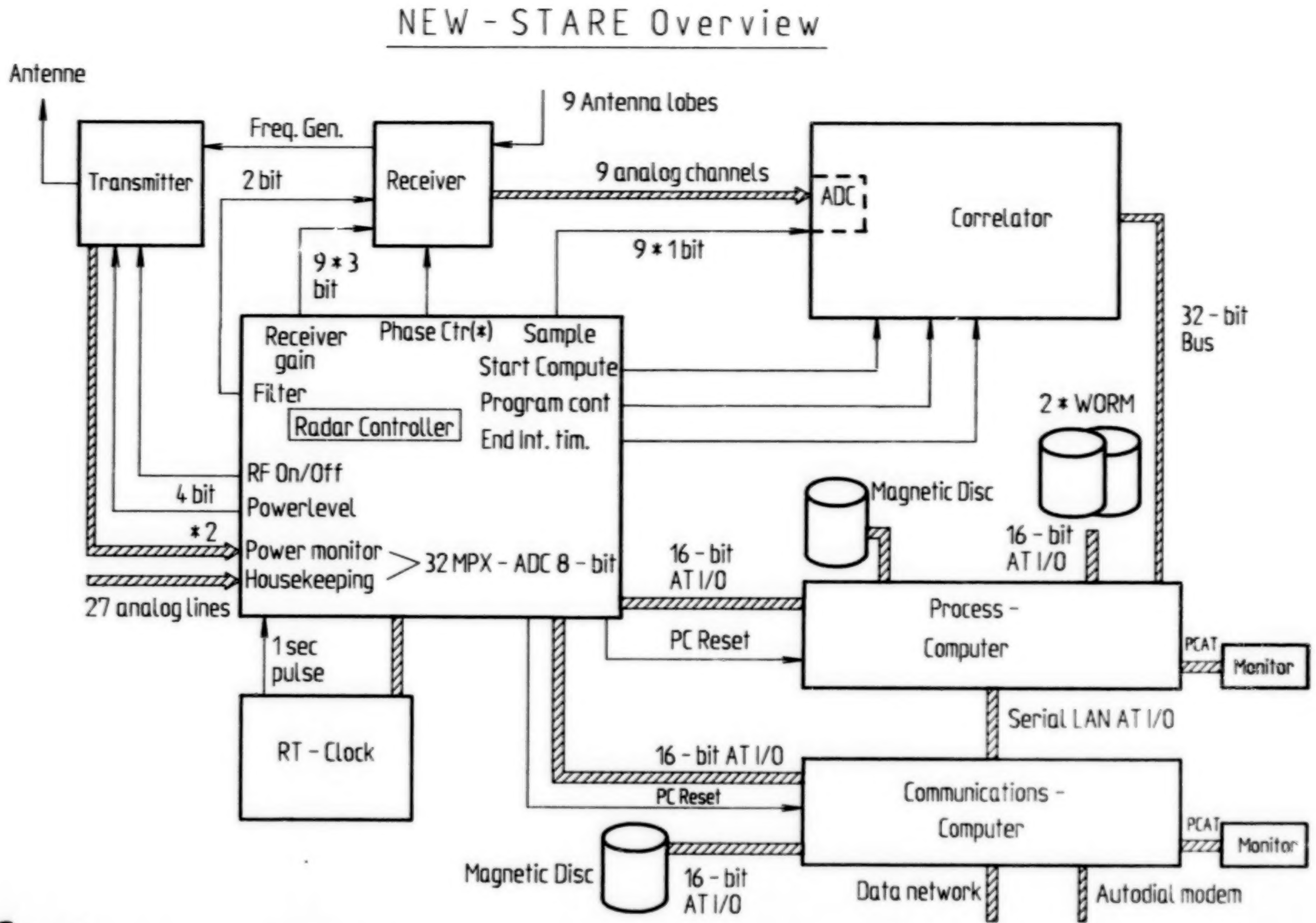


Figure 8: Simple block diagram of a New Stare radar station (see text).

to optimize the receivers for the following integration time. The time sequence during one scan are determined by the content of two memories, called 'the pattern' - and 'the dwell time' memory, respectively. The operations these memories control are: the radio frequency on/off (the transmission of pulses), the phase controll of the transmitted signal, the sample times, and at the end of a scan a 'start compute' signal (or in the case an alternating code is used a 'continue' signal) is send to the correlator. Each 16 bit word in the pattern memory determines the setting of the radar controller outputs, except the power level and the end of integration time. Each bit in the word controlls one output. To each pattern-word there is a corresponding word in the (parallel) dwelltime memory, which defines the time interval during which the settings given by the pattern-word must remain at the outputs. At the end of that time interval the next words are processed, and so on untill the end of the scan. The scan is repeated a certain number of times in each integration interval.

The correlator is responsible for digitizing the coherently detected signals on the outputs of the 9 receivers, under control of the sample pulses from the radar controller. The data are stored in the data memory. The processer part carry out the complex multiplications of two complex input values, using the content of a program memory (loaded with the file generated in the process computer), which determines the addresses of the data to be multiplied. The program memory also determines the addresses in the result memory of these relevant correlation marix elements, which are then added to and stored in the result memory. The correlator is designed to handle multi pulse patterns as well as pulse compression.

The communications computer is with a local area network in contact with the process computer and its devices. This computer is used partly to display observations in real time at the radar station, and partly to facilitate remote access to the station for control of experiments or for transmission of data to an experimenter, who may for example be located at a rocket range, where real time data can be used in the decision process to launch rockets [Nielsen, 1983b].

6.2 Antenna

Consider a group of n similar antennas, which are oriented the same way as a reference antenna. The location of the j'th antenna is given by the vector $\mathbf{r}_j$. The amplitude and phase of the current in the j'th antenna is given in terms of the current, I, in the reference antenna

$$I_j = \sigma_j I \tag{47}$$

Let the far field antenna pattern of the reference antenna be

$$\psi_O = Ie^{i\frac{kR}{R}} \cdot F(\theta, \varphi) \tag{48}$$

where $k = \frac{2\pi}{\lambda}$, R is the distance to the point of observation, and F has its maximum value in the direction of maximum gain. Let φ be measured in the xy-plane from the x-axis, and θ from the z-axis. The antenna field of the antenna array consisting of all n antennas is

$$\psi = \sum_{j=1}^{n} I_j e^{\frac{ik \cdot R_j}{R_j}} F(\theta, \varphi) \tag{49}$$

and using the approximation $\underline{R}_j = \underline{R} + \underline{r}_j \cdot \hat{\underline{R}}$

$$\begin{aligned} \psi &= \psi_o \sum_{j=1}^{w} \sigma_j e^{-ik\underline{r}_j \cdot \hat{\underline{R}}} \\ &= \psi_o g(\theta, \varphi) \end{aligned} \tag{50}$$

so the field of the array is given by the product of the field of the reference antenna and the gain function g(θ, φ). Let the referene antenna be at the origo of a coordinate system, where the j'th antenna has the coordinates (x_j, y_j, z_j). The gain function can then be written

$$g(\theta, \varphi) = \sum_{j=1}^{w} \sigma_j e^{ik(x_j sin\theta cos\varphi + y_j sin\theta sin\varphi + z_j cos\theta)} \tag{51}$$

and the magnitude of the gain function yields the radiation pattern, a function of θ and ϕ,

$$|g| = \sqrt{Re^2(gg*) + Im^2(gg*)} \tag{52}$$

Several coherent radars make use of an antenna arrangement producing a single narrow antenna beam, which is used both for transmitting and receiving [for example Providakes et al., 1983]. This technique has its major advantage in a good suppression of the sidelobes. In this special case all the antennas may be located on the x-axis with a constant distance, d_x, a constant amplitude, I/n, and a constant phase change, δ_x, between all neighboring antennas, thus

$$g_x(\theta, \varphi) = \sum_{j=1}^{n} \frac{1}{n} e^{-i(j-1)(\delta_x + kd_x sin\theta cos\varphi)} \tag{53}$$

In this case it is easy to derive $\|g_x\|$, and introducing $\alpha = \delta_x + kd_x sin\theta cos\varphi$, we find that

$$|g_x(\theta, \varphi)| = \frac{sin\frac{n\alpha}{2}}{nsin\frac{\alpha}{2}} \tag{54}$$

For antennas distributed along the y- or z-axis expressions equivalent to Equation 54 for the associated gain functions can be found.

A single string of antennas will have a gain function which has essentially the form of a disk perpendicular to the direction of the string. To make the array more directional in, say, the direction of the y-axis two or more strings of antennas could be located at larger z-values and parrallel to the x-axis.

The above expressions are valid for antennas in free space. If the antennas are located over conducting ground, the mirror image of the antennas in the ground must be considered by setting (with the z-axis perpendicular to the ground) $n_z=2$, $\delta_z = \pi$, and $d_z = 2$ h, where h is the height of the antennas over the ground; the radiation pattern owing to the ground is therefore modified by

$$|g_g(\theta,\varphi)| = sin\left(\frac{2\pi hcos\theta}{\lambda}\right) \tag{55}$$

and the resulting radiation function of the string of antennas along the x-axis, also considering the radiation pattern of the reference antenna, has the form

$$|g'_x| = F(\theta,\varphi) \quad |g_x| \cdot |g_g| \tag{56}$$

The maximum gain in dB of an array with n antennas is to a reasonable approximation given by: 10 log[n]+(gain of reference antenna). The maximum gain can also be expressed by the 3 dB beamwidth's $\Delta\theta$ and $\Delta\varphi$, measured in radians, as

$$G = \frac{4\pi}{\Delta\theta\Delta\varphi} \tag{57}$$

and the effective area, A_e, is

$$A_e = \frac{\lambda^2}{\Delta\theta\Delta\varphi} \tag{58}$$

To lower the sidelobe level one can use tapering, i.e. the gradual reduction of the current amplitude going from the center towards the sides of the array. This 'spatial' variation of the current amplitude could for example have the form of a triangle, of a Chebyshevs polynomium or follow a binominal law; in the latter case the sidelobes actually disappear. Tapering has in addition the unwanted effect of broadening the main antenna lobe, and therefore reducing the gain and effective area of the array.

A much used technique for constructing a string of dipole antennas, which is leightweight, portable, easy to construct and to set-up, and also fairly inexpensive, is the co-co array (coaxial-colinear array). It is constructed of coaxial cable in which the inner and outer conductor are interchanged every half wavelength, thereby producing a string of colinear halfwave dipoles. The antenna is fed at the junction between the two center dipoles. The feed arrangement consists of an impedance matching network and a balun [Balsley and Ecklund, 1972]. In such an array the current amplitude in the dipoles decreases away from the feed point owing to cable losses; thus, tapering can to some degree be controlled by choise of cable.

Antenna arrays with a single narrow beam are primarily used in coherent radars dedicated to plasma physics studies. Radar systems used to measure the spatial patterns of electron drift velocities in the high latitude ionosphere uses antenna systems with several narrow antenna lobes allowing observations to be made over a large area with good directional resolution. For example, the two Stare radar stations are located such that the lines of sight of their antenna beams cross each other in regions of good aspect angle. Thus, backscatter from the cross-over regions can be measured from two directions simultaneously. This capability gave rise to important new possibilities for plasma physics studies, and is the property of the system that allows estimation of the electron drifts in the ionosphere.

The receiver array of the Stare system consists of 16 masts each with a stack of four yagis separated by one wavelength. The distance between neighbouring masts is one wavelength. The four yagis are combined and connected to a pre-amplifier. The output of the 16 pre-amplifiers are connected to a Butler matrix, which is a passive phasing network that forms, simultaneously, 16 narrow beams each with its separate output. Each of the narrow beams is determined by the following expression for α (Equation 54)

$$\alpha = (2m-1)\frac{\pi}{16}(n_z - 1) + d_z sin\theta cos\varphi \tag{59}$$

where $d_z = \lambda$, n_z varies from 1 to 16, and for beams 1 to 16 m varies from -7 to + 8. The property of the Butler matrix is such as to produce, simultaneously, the changes in phase from one mast to the next and for each beam, as given by the first term on the RHS. Inserting numerical values we find the beamwidth to be 3.2° and the separation between beams to be 3.6° (the lobes cross at the 4 dB points). The transmitter antenna is wide beam so that the transmitted signal propagates simultaneously along the receiver antenna lobes. The most important feature of the transmit/receive arrays is that, without loss of gain, one can measure in 16 independent directions simultaneously. This allows observations in two dimensions (essentially latitude and longitude) to be obtained with much better time resolution than would be possible using an electronically switched phased-array antennas. The main disadvantage is the relative poor sidelobe supression compared to that which can be obtained with phase-arrays, i.e. using the same narrow beam antenna for both transmit and receive. Note that the many lobes in the radiation pattern can also be formed by applying a software-simulation of a Butler matrix to the signals from the 16 masts [McNamare et al., 1983]. One advantage of this is the ease with which one can move all 16 antenna lobes in azimuth, and thus essentially change the field of view. The main limitation on such motions is the requirement of good aspect angle in the field of view.

At lower frequencies, in the HF frequency band (for example from 8 to 20 MHz), one is forced to use phased-arrays [Greenwald et al., 1984]. The principal reason is that a Butler matrix only can be designed for one frequency. In the HF band, however, one must be able to select the frequency which ensures good aspect angle conditions in the ionosphere, i.e. select the signal frequency which for the given electron density distribution in the ionosphere bends the signal path to make it perpendicular to the geomagnetic field lines in the region with irregularities such that backscatter occurs. Another reason is, that the strong possibility of interference from man-made noise makes it necessary to be able to select the operating frequency.

7. Electron drift velocities.

For the study of solar-wind/magnetosphere/ionosphere interactions it is of fundamental importance to have information on the horizontal ionospheric plasma flows. A unique property of coherent radar systems is their ability to provide such information. The key point here is that the mean Doppler velocity of the backscattered radar signal contains information about, is a function of, the DC electric field in the backscatter region [Nielsen and Whitehead, 1983; Nielsen et al., 1983b; Nielsen and Schlegel, 1983, 1985; and Reinleitner and Nielsen, 1985]. With observations in a given backscatter volumen of the Doppler velocity from two (or more) directions, these directional velocities allow an estimate of the electric field in that volume. In the same sense, if directionally independent observations of the Doppler velocity are made in different backscatter volumens, then they can be combined to yield an estimate of the electric field if it can be argued that the field is the same in these volumes.

Figure 9 shows the horizontal radiation pattern of the Stare radars receiving arrays. The set up allows Doppler velocities to be measured from two directions in several backscatter volumes. Each pair of Doppler velocities can now be combined in one of two ways to yield an estimate of the electron drift velocity, or, since the electrons are ExB drifting, the electric field. In the first approximation Equation 5 states that the Doppler velocity in a given direction equals the component of the electron drift velocity on that direction. With two components of the electron drift the total drift velocity is determined. This estimate of the electron drift velocity turns out to yield a good estimate of the actual flow direction, but to underestimate the actual magnitude of the electron speed - the more so the larger the electron speed. However, it is important that the estimate is a lower one, it will not be some wild big number. It is a conservative estimate. The observations shown in Figure 3, panels a and b, confirm that the Doppler velocities measured at large angles to the direction of the electron flow are good approximations of the electron velocity components, but also that at smaller angles the magnitude of the Doppler velocity is primarily a measure of the electron speed (Figure 3b). With the direction of the electron flow known from the first approximation, one can now select the Doppler velocity associated with a small flow angle to determine the total flow speed, and combining this with the Doppler velocity measurement at large flow angle yields an electron drift velocity estimate, which agrees very well in direction with the actual drift velocity, and quite well (usually better than a factor two) with its magnitude.

Figure 10 shows an example of the kind of observations a VHF coherent radar system can provide. Flows in the large field of view with the good spatial resolution is measured, typically, every 20 seconds. The data allows the spatial variations in the flow to be determined at a given time, and to follow how this flow changes as a function of time. This is the unique feature of a coherent radar system.

Returning to the estimation of electron drift velocities. If only one radar station with a multi-lobe receiver array with a radiation pattern as in Figure 9 is available, an electron drift velocity can still be estimated. If the electron flow is constant along a curve (say a constant geomagnetic latitude contour) through the array lobes, and if all Doppler velocities at the given contour are measured at large flow angles in all the lobes (Equation 5 valid), then these Doppler velocities can be combined to yield an estimate of the electron flow at that contour. This is illustrated in Figure 3a, where the measurements at large flow angles are fitted in a least square procedure to a circle. The diameter of the circle (through the origo) is a measure of the electron drift velocity (as expressed in Equation 5). This technique has been widely used in single stations HF radars, and it has been shown that

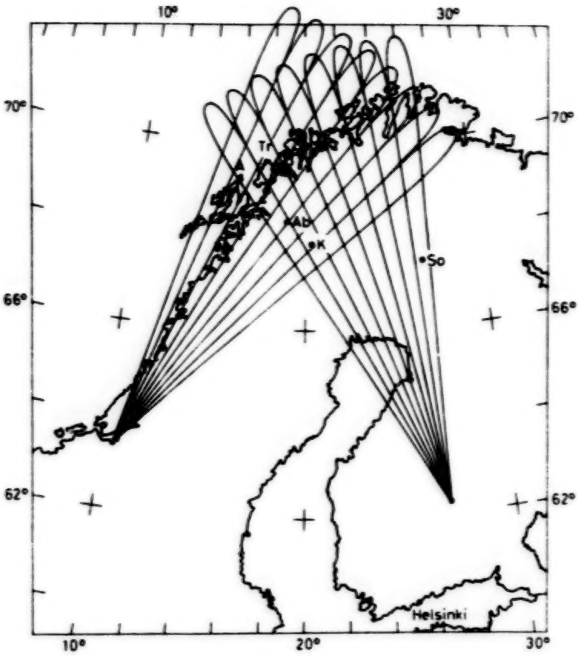


Figure 9: Map of Scandinavia showing the eight central lobes of the receiver array radiation patterns from the two Stare radar stations (from Greenwald et al. [1978])

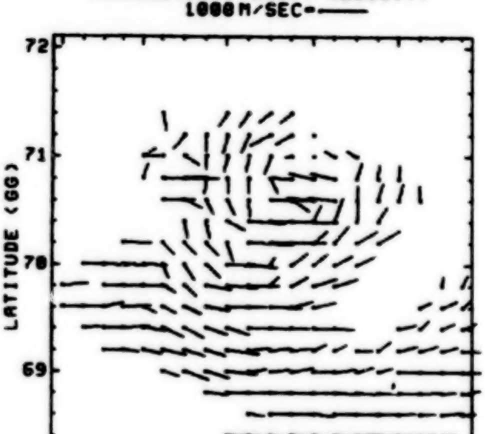


Figure 10: A typical example of drift velocity measurements made with the Stare radar system (from Nielsen [1982]).

even in cases where the flow on the contour is not spatially homogeneous the technique can nevertheless yield valuable information also on the spatial variations [Ruohoniemi et al., 1989].

8. Conclusion

Coherent radars are at present very actively used research tools. Use of these radars to study the physics of ionospheric irregularities continues. Simultaneous measurements at multiple frequencies have been made and more are planned; joint measurements with incoherent - and coherent radars will yield more detailed information about the backscattering medium than previously available, and that can also be achieved using rocket born instruments for in situ measurements. Multi-lobed array radiation patterns have so far only been used at high latitudes. They should also be employed in other parts of the Earth's ionosphere to determine the spatial variations of the radar signals and their evolution with time. The first radar system designed for electron drift velocity measurements was the Stare system, which started operations in 1977. It has been followed by similar systems, by Sabre [Nielsen et al., 1983a] and by Bars [McNamara et al., 1983]. Also at lower frequencies (HF) radars are used for geophysical studies [Greenwald et al., 1985; Dudeney, 1988]. It has been noted that a coherent radar on Kergulen Island, in the southern hemisphere, would have its field of view in the region magnetic conjugate to the Stare field of view in the northern hemisphere. With the large field of view and good spatial resolution of these radar systems detailed studies of ionospheric flows in conjugate regions could be made. Such data are at present available only from coherent HF radars with their fields of view inside the polar caps.

9. References

Andre, D. , The Dependence of the relative backscatter cross section of 1-m density fluctuations in the auroral electrojet on the angle between electron drift and radar wave vector, *J. Geophys. Res.*, **88,** 8043, 1983.

Blasley, B.B. , and D.T. Farley, Radar studies of the equatorial electrojet at three frequencies, *J. Geophys. Res.*, **76,** 8341, 1971.

Balsley, B.B. , and W.L. Ecklund, VHF power spectra of the radar aurora, *J. Geophys. Res.*, **77,** 4746, 1972.

Balsley, B.B. , G. Haerendel, and R.A. Greenwald, Equatorial spread F: Recent observations and a new interpretation, *J. Geophys. Res.*, **77,** 5625, 1972.

Balsley, B.B. , and W.L.Ecklund, A Portable coaxial collinear antenna, IEEE Transaction on Antennas and Propagation, 1972.

Carter, D.A. , B.B. Balsley, and W.L. Ecklund, VHF Doppler radar observations of the African equatorial electrojet, *J. Geophys. Res.*, **81,** 2786, 1976.

Cohen, R. , and K.L. Bowles, Secondary irregularities in the equatorial electrojet, *J. Geophys. Res.,* **72,** 822, 1967.

Dudeney, J.R. , The Polar Anglo-American Conjugate Experiment, *EOS,* **69,** 1347, SA31A-04, 1988.

Ecklund, W.L. , D.A. Carter, and B.B. Balsley, Gradient drift irregularities in mid-latitude sporadic E, *J. Geophys. Res.,* **86,** 858, 1981.

Farley, D.T. , H.M. Ierkic, and B.G. Fejer, Radar interferometry: A new technique for studying plasma turbulence of the ionosphere, *J. Geophys. Res.,* **86**, 1467, 1981.

Farley, D.T. , Multiple-pulse incoherent scatter correlation function measurements, *Radio Sci.,* **7,** 661, 1972.

Farley, D.T. , Incoherent scatter correlation function measurements, *Radio Sci.,* **4,** 935, 1969.

Fejer, B.G. ., D.T. Farley, B.B. Balsley, and R.F. Woodman, Radar observations of two-dimensional turbulence in the equatorial electrojet, 2. *J. Geophys. Res.,* **81,** 130, 1976.

Fejer, B.G. , J. Providakes, and D.T. Farley, Theory of plasma waves in the auroral E region, *J. Geophys. Res.,* **89,** 7487, 1984.

Fejer, B.G. , and M.C. Kelley, Ionospheric irregularities, *Rev. of Geophys. and Space Phys.,* **18,** 401, 1980.

Greenwald, R.A. , and W.L. Ecklung, A new look at radar auroral motions, *J. Geophys. Res.,* **80,** 3642, 1975.

Greenwald, R.A. , W. Weiss, E. Nielsen, and N.R. Thomson, STARE: A new radar auroral nackscatter experiment in northern Scandinavia, *Radio Sci.,* **13,** 1021, 1978.

Greenwald, R.A. , K.B. Baker, R.A. Hutchins, and C. Hanuise, An HF phased-array radar for studying small-scale structure in the high-latitude ionosphere, *Radio Sci.,* **20,** 63, 1984

Haldoupis, C. , and G. Sofko, Doppler spectrum of 42 MHz CW auroral backscatter, *Can. J. Phys.,* **54,** 1571, 1976.

Haldoupis, C. , and E. Nielsen, 140 MHz auroral backscatter: Evidence for characteristic phase velocities below and above the ion acoustic speed, *Geophys. Res. Lett.,* in the press, 1989.

Harang, L. , and W. Stoffregen, Echoversuche auf Ultrakurzwellen, Hochfrequenztechnik und Elektroakustik, Akademische Verlagsgesellschaft m.b.H., Leipzig, 55, 105-108, 1940.

Ierkic, H.M. , Radar observations of the equatorial electrojet irregularities and theory of type 1 turbulence, *Ph. D. Thesis,* Cornell Uni,N.Y., 1980.

Keys, J.C. , and M.K. Andrews, Gravity wave and sporadic E echo signatures on VHF backscatter radar systems, *Planet.Space Sci.*, **32,** 1455, 1984.

Kudeki, E. , D.T. Farley, and B.G. Fejer, Long wavelength irregularities in the equatorial electrojet, *Geophys. Res. Lett.*, **9,** 684, 1982.

Lehtinen, M.S. , and I. Haggstrom, A new modulation principle for incoherent scatter measurements, *Radio Sci.*, **4,** 625, 1987

Moorcroft, D.R. , Estimates of absolute scattering coefficients of radar aurora, *J. Geophys. Res.*, **92,** 8723, 1987.

McNamara, A.G. , D.R. McDiarmid, G.J. Sofko, J.A. Koehler, P.A. Forsyth, and D.R. Moorcroft, BARS- A dual bistatic auroral radar system for the study of electric fields in the Canadian sector of the auroral zone, *Adv. Space Res.*, **2,** 145, 1983.

Nielsen, E. , M. Uspensky, A. Kustov, A. Huuskonen, and J. Kangas, On the dependence of the Farley-Buneman turbulence level on the ionospheric electric field, *J. Atmos. Terr. Phys.*, **50,** 601, 1988.

Nielsen, E. , Aspect angle dependence of backscatter intensity of 1-m auroral plasma waves, *J. Geophys. Res.*, **93,** 4119, 1988.

Nielsen, E. , Aspect angle dependence of mean Doppler velocities of 1-m auroral plasma waves, *J. Geophys. Res.*, **91,** 10173, 1986.

Nielsen, E. , and K. Schlegel, A first comparison of STARE and EISCAT electron drift velocity measurements, *J. Geophys. Res.*, **88,** 5745, 1983.

Nielsen, E. , and K. Schlegel, Coherent radar Doppler measurements and their relationship to the ionospheric electron drift velocity, *J. Geophys. Res.*, **90,** 3498, 1985.

Nielsen. E. , C.I. Haldoupis, B.G. Fejer, and H.M. Ierkic, Dependence of auroral power spectra variations upon electron drift velocity in the eastward electrojet, *J. Geophys. Res.*, **89,** 253, 1984.

Nielsen, E. , W. Guettler, E.C. Thomas, C.F. Stewart, T.B. Jones, and A. Hedberg, A new radar auroral backscatter experiment, *Nature*, **304,** 712, 1983a.

Nielsen, E. , J.D. Whitehead, L.A. Hedberg, and T.B. Jones, A Test of the cosine relationship using three-radar velocity measurements, *Radio Sci.*, **18,** 230, 1983b.

Nielsen, E. , and J.D. Whitehead, Radar auroral observations and ionospheric electric fields, *Adv. Space Res.*, **2,** 131, 1983.

Nielsen, E. , Real-time ionosphere data, *EOS*, **64,** 596, 1983.

Nielsen, E. , The Stare system and some of its applications, The IMS Source Book, AGU, 1982.

Prakash, S. , C.L. Jain, B.B. Balsley, and R.A. Greenwald, Evidence of two types of electron density irregularities in the electrojet over Thumba, India, *J. Geophys. Res.*, **79**, 4334, 1974.

Providakes, J.F. , Radar interferometer observations and theory of plasma irregularities in the auroral ionosphere, *Ph. D. Thesis,* Cornell Uni.,N.Y.,1985.

Providakes, J.F. , W.E. Swartz, D.T. Farley, B.G. Fejer, First VHF auroral radar interferometer observations, *Geophys. Res. Lett.*, **10**, 401, 1983.

Reinleitner, L.A. , and E. Nielsen, Self-consistent analysis of electron drift velocity measurements with the Stare/Sabre system, *J. Geophys. Res.*, **90**, 8477, 1985.

Riggin, D. , W.E. Swartz, J. Providakes, and D.T. Farley, Radar studies of long-wavelength waves associated with mid-latitude sporadic E layers, *J. Geophys. Res.*, **91**, 8011, 1986.

Rummler, W.D. , Two pulse spectral measurements, Tech. Memo. MM-68-4121-15, Bell Telephone Lab., Whiffany, N.J., 1968.

Ruohoniemi, J.M. , R.A. Greenwald, K.B. Baker, J.-P. Villain, C. Hanuise, and J. Kelly, *J. Geophys. Res.*, in press, 1989.

Schmidt, W. , R. Pellinen, J. Silen, and E. Nielsen, New Stare, Internal Report Finnish Meteorological Institute, 1988.

Tsunoda, R.T. , R.I. Presnell, and R.L. Leadabrand, Radar auroral echo characteristics as seen by a 398-MHz phased array radar operated at Homer, Alaska, *J. Geophys. Res.*, **79**, 4709, 1974.

Turunen, T. , and J. Silen, Modulation patterns for EISCAT incoherent scatter radar, *J. Atmos. Terr. Phys.*, **46**, 593, 1984.

Unwin, R. , and W.J. Baggaley, The radio aurora, *Ann. Geophys.*, **28**, 111, 1972.

Walker, A.D.M. , R.A. Greenwald, and K.B. Baker, Determination of the fluctuation level of ionospheric irregularities from radar backscatter measurements, *Radio Sci.*, **22**, 689, 1987.

Whitehead, J.D. , On reducing the inherent noise in the double-pulse Doppler technique, *Radio Sci.*, **18**, 93, 1983.

Chapter 12

REMOTE SENSING OF THE HIGH-LATITUDE IONOSPHERE WITH COHERENT SCATTER HF RADAR: A REVIEW OF THE TECHNIQUE AND RECENT WORK

J. M. Ruohoniemi, R. A. Greenwald, and K. B. Baker

The Johns Hopkins University Applied Physics Laboratory
Laurel, Maryland 20707

ABSTRACT

The high-latitude ionosphere is being studied with a new type of HF radar. Facilities are now in operation at Goose Bay, Labrador, Schefferville, Quebec, and Halley Station, Antarctica. In this paper, we review the HF coherent scatter technique for the remote sensing of ionospheric structure and dynamics. We describe the operation of the Goose Bay radar and the real-time control of operating parameters that can be exercised during campaign periods. We illustrate the methods for mapping ionospheric plasma convection. We discuss recent findings on the generation of small-scale (~ 10 m) irregularities in the E region by field-aligned currents and the occurrence of irregularities in the wintertime F region trough at dusk. The dependence of plasma convection in the dayside ionosphere on the IMF and the conjugate relationship is illustrated with a case study involving the Goose Bay and Halley Station radars. We also review recent work showing that the passage of gravity waves in the neutral atmosphere is detected, extending the scope of the HF radar studies to important questions of ionosphere-thermosphere coupling.

INTRODUCTION

The high-latitude regions present unique opportunities for the study of a wide range of problems in basic plasma physics and the interaction of the solar wind with the earth's magnetosphere. It is known, for example, that plasma instabilities are triggered by plasma drifts and density structures that arise within the high-latitude regions (*for a review, see Keskinen and Ossakow* [1983]). Also, large-scale electric fields generated within the outer magnetosphere map to the ionosphere, primarily in the high-latitude zone. Of particular interest to WITS, the interaction of the active ionosphere of the auroral regions with the thermosphere is believed to be a major source of energy for the neutral atmosphere.

In this paper, we describe a technique developed in this decade for the study of high-latitude electrodynamics. It is based on the refraction and scattering of high frequency (HF) radio signals within the ionosphere. The experimental facilities, of which the Johns Hopkins University/Applied Physics Laboratory (JHU/APL) radar located at Goose Bay, Labrador is the best-known example, utilize electronic beam forming and steering for an exceptional combination of spatial coverage (~ 10^6 km), spatial resolution (~ 50x50 km), and temporal resolution (~ 1½ min). The HF radars at Schefferville, Quebec and Halley Station, Antarctica have more recently become operational. The Schefferville radar has a field of view that largely overlaps that of the Goose Bay radar and the field of view of the Halley Station radar is magnetically conjugate to that of Goose Bay. The use of the radars for studying conjugate phenomena is referred to as the Polar Anglo-American Conjugate Experiment (PACE). We illustrate with examples the range of studies that can be carried out. In particular, the drift of ionospheric plasma induced by electric fields is readily observed and can often be extensively mapped. With this information, the drift conditions for the occurrence of irregularities can be analyzed. We present the results of recent work showing the relation of particular types of irregularity activity to ionospheric plasma drift. Of course, the mapping of high-latitude plasma convection is itself an important research goal. We show examples of convection maps and discuss the dependence of convection in the dayside ionosphere on the IMF. We conclude with a review of recent work on the effects of gravity waves on HF propagation and show that the Goose Bay radar in fact observes strong gravity wave disturbances.

HF Radio Wave Propagation

In the past, most coherent scatter studies of the ionosphere were performed with very high frequency (VHF) 30-300 MHz and ultrahigh frequency (UHF) 300-3000 MHz radars, which are essentially unaffected by changes in ionospheric properties. However, it has long been known that coherent scatter can also be obtained with high frequency (HF) 3-30 MHz radars [*Bates and Albee*, 1970]. At these lower frequencies the ionosphere has marked effects on the propagation of radar signals.

Figure 1 illustrates the manner in which VHF and HF radar signals are scattered by small-scale irregularity structures in the high-latitude E and F regions. The irregularities responsible for coherent scatter are aligned with the lines of force of the earth's magnetic field (*see Fejer and Kelley* [1980]). Signals from a VHF radar propagate in straight lines. If they encounter irregularities along their path, some of the energy in the incident wave will be scattered. If the scattered signals are to return to the radar, the incident wave must be directed normal to the irregularities in the region of interest and hence normal to the magnetic field. At very high latitudes, the direction of propagation of the probing radar signal must become nearly horizontal. This condition can be satisfied by a VHF radar in the E region but not in the F region. At HF frequencies the radar signal is refracted by the ionosphere toward the horizontal and the orthogonality condition can be satisfied in both the E and F regions.

It is necessary but not sufficient that the radar signal propagate orthogonal to **B**. Density structure of a scale that matches the radar sampling wavelength must also be present. The intensity of the backscattered signal is greatly increased if reflections from successive phase fronts of a density perturbation constructively interfere. This will occur if the scale length of the irregularity is half that of the radar wavelength. Thus, HF radars operating in the 8-20 MHz range of frequencies are sensitive to irregularities of scale length 7.5-19 m.

Coherent scatter radars are similar to incoherent scatter radars in that they sense density perturbations in the plasma density distribution. However, the perturbations responsible for incoherent scatter are generated by thermal fluctuations within the plasma, while the irregularities responsible for coherent scatter are amplified well above thermal levels by the action of plasma instabilities. Coherent scatter consequently does not require the high radar power levels associated with incoherent scatter, but the ability to make measurements depends more on the presence of irregularities within the plasma medium.

It will be clear from the preceding discussion that the successful operation of a coherent scatter HF radar requires an appropriate choice of transmitting frequency. This will need to be varied throughout the day as ionospheric conditions change, and also in accordance with solar cycle variations and storm-related ionospheric disturbances. Also, because complex modes of propagation are possible, it should be possible to vary the range interval for sampling returns. Thus the operating parameters of a coherent scatter HF radar must be kept flexible.

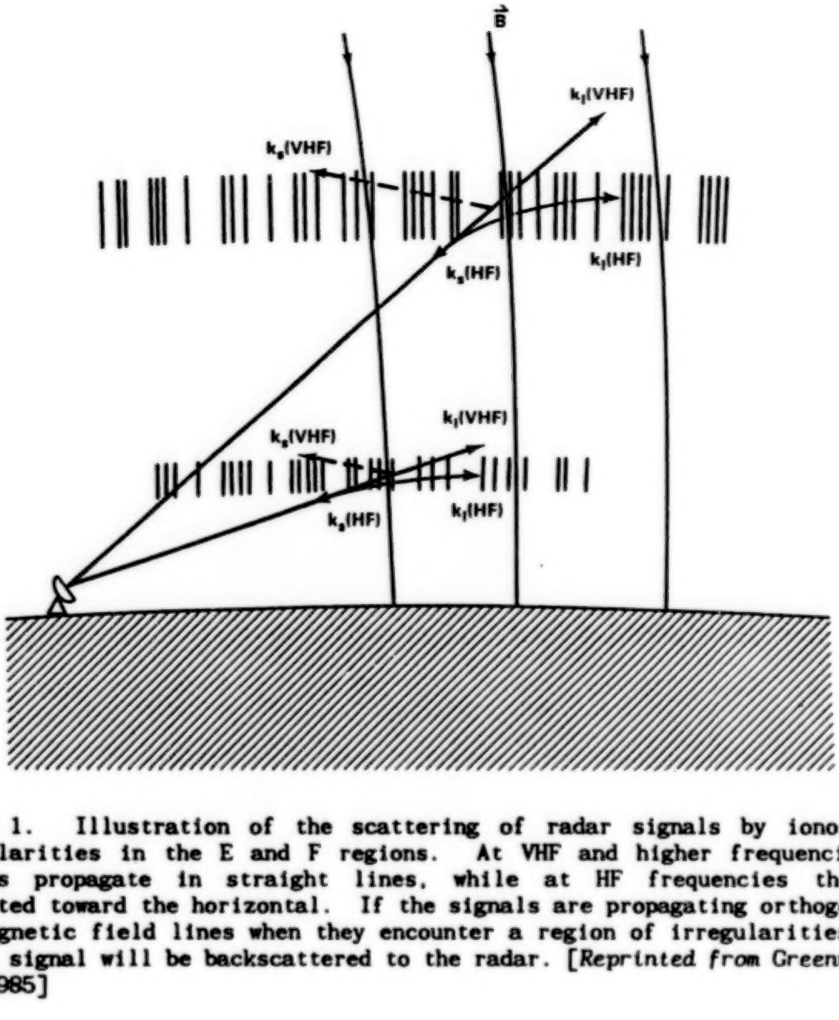


Figure 1. Illustration of the scattering of radar signals by ionospheric irregularities in the E and F regions. At VHF and higher frequencies the signals propagate in straight lines, while at HF frequencies they are refracted toward the horizontal. If the signals are propagating orthogonal to the magnetic field lines when they encounter a region of irregularities, some of the signal will be backscattered to the radar. [*Reprinted from Greenwald et al.*, 1985]

In the next section, we briefly discuss the Goose Bay HF radar. The sister facilities at Schefferville, Quebec (Centre National de la Recherche Scientifique, France) and Halley Station, Antarctica (British Antarctic Survey) are basically similar. For a more detailed dicussion of the HF radar operations, see *Greenwald et al.* [1985].

HF Coherent Scatter Radar

The JHU/APL HF radar at Goose Bay, Labrador (53.4°N, 60.4°W) looks over northeastern Canada and Greenland (Figure 2). The radar consists of a pair of 16-element electronically steered arrays, where each element is a broadband (8-20 MHz) log periodic antenna. The arrays are 240 m long, 15 m above the ground, and separated by 100 m along the direction of the common array normal, which is directed 5° to the east of north. The narrow (~ 4°) antenna beam is scanned in a clockwise direction over a 50° azimuth sector through 16 evenly spaced pointing azimuths. The usual integration time for each beam position is 5-6 s and thus a single scan requires only 80-96 s. When the radar acts as vertical interferometer to measure the elevation angle of arrival by receiving on the two arrays simultaneously, the integration time is doubled. For each pointing azimuth, the returns from 50 range gates are sampled simultaneously. The gate separation is usually 30 or 45 km. The range to the first gate can be set to any value, but is generally close to 300 or 600 km.

The electronic steering of the antenna is accomplished with the broadband phasing matrix shown schematically in Figure 3. The principal elements of the phasing matrix are the eight branching networks known as 'phasing trees'. On transmission, signals enter a tree from the receiver side and pass through a network of power dividers and selected lengths of coaxial cable that ultimately produces successive time delays among the signals at the outputs. On receive, signals enter the tree from the antenna side and are reformed into a composite signal. Those signals arriving from the direction into which the antenna is directed will undergo constructive interference and exit the array at a considerably enhanced power level. The delay lines produce constant time delays rather than phase delays through the phasing matrix, and therefore the direction of the antenna beam is independent of frequency.

Each phasing tree produces one possible set of time delays across the antenna array. The 4-way switches shown in Figure 3 are used to select one of the phasing trees, and the double-pole double-throw (crossover) switches are used to determine whether the selected beam is directed to the right or left of the array normal. The properly phased signals are then directed to the T/R switch and chain of amplifier modules at the base of each antenna. The pulsed power level from each amplifier chain varies between 500 W and 1000 W, depending on frequency. (A modification to the transmitters will soon render the power level independent of frequency.) While the average peak radiated power from the array is thus only 10 kW, the array factors and antenna gain result in an effective radiated power of nearly 1 MW.

The radar transmits the seven-pulse pattern shown in Figure 4. An on-site microcomputer samples the backscattered signals from this pattern and calculates, in real-time, 17-lag complex autocorrelation functions (0 to 16 t_0). A seven-pulse pattern is sufficient to yield all the lags for the autocorrelation analysis. The autocorrelation data for the twenty range gates that return the strongest signal are stored on magnetic tape for further processing at APL.

The microcomputer at the station controls the radar as well as performs the preliminary processing of the incoming data. Two modes of operation are possible. In one, an operator who may be on-site or at a distant location equipped with a modem, enters operating parameters into the microcomputer. This mode is used for highly interactive real-time operation of the radar,

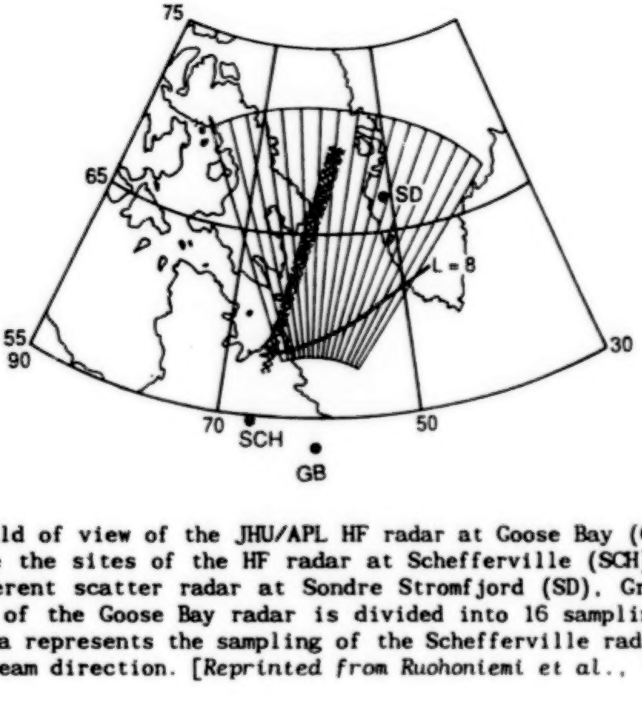


Figure 2. Field of view of the JHU/APL HF radar at Goose Bay (GB), Labrador. Also shown are the sites of the HF radar at Schefferville (SCH), Quebec, and the SRI incoherent scatter radar at Sondre Stromfjord (SD), Greenland. The field of view of the Goose Bay radar is divided into 16 sampling directions. The shaded area represents the sampling of the Schefferville radar on a single intersecting beam direction. [*Reprinted from Ruohoniemi et al.*, 1989]

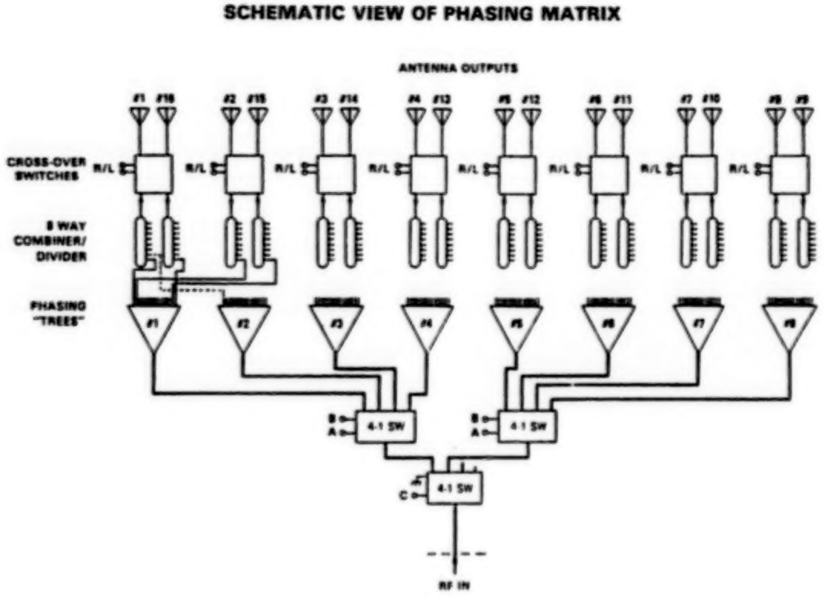
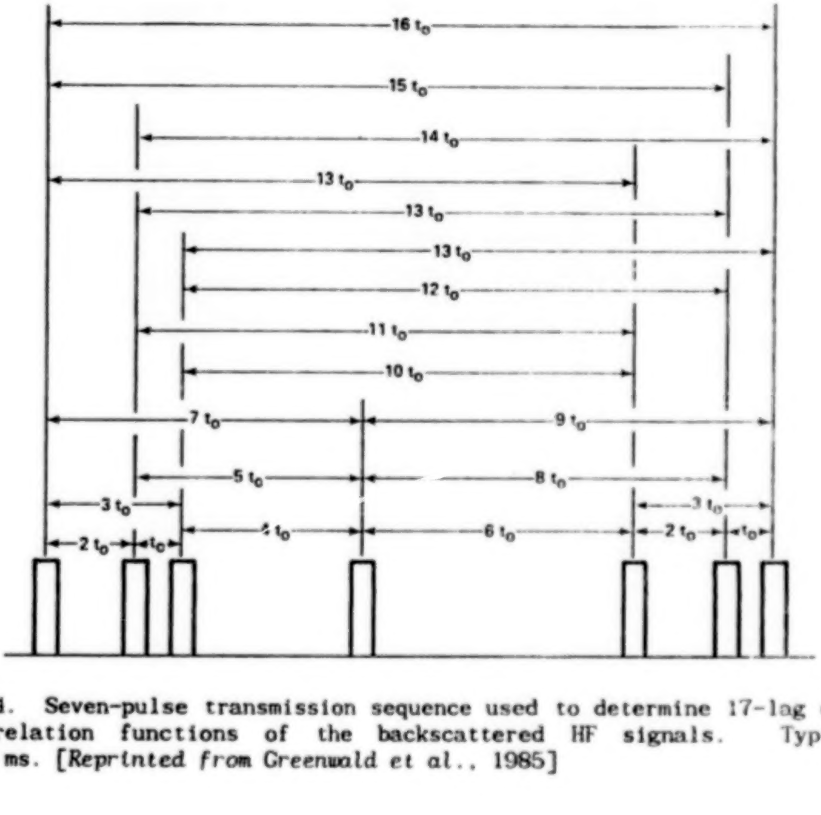


Figure 3. Schematic diagram of the phasing matrix of the JHU/APL HF radar. A progressive time delay is established across the antenna array on transmission and reforms the beam on reception. [*Reprinted from Greenwald et al.*, 1985]



7-PULSE PATTERN FOR 17-LAG AUTOCORRELATION FUNCTION

Figure 4. Seven-pulse transmission sequence used to determine 17-lag complex autocorrelation functions of the backscattered HF signals. Typically, $t_0 = 3.9$ ms. [*Reprinted from* Greenwald *et al.*, 1985]

such as would be desirable during campaign periods. The other mode is used for noninteractive operation and consists of command strings that are stored on the disk in the form of text files. The strings determine the operating times, modes, and frequencies of the radar. They may be tailored to a particular type of study and supplied weeks or months in advance.

The Goose Bay radar began operation in 1983 and has run continuously (24-hr, 365 days/yr) since 1985 with occassional interruptions for system upgrades. The facility at Schefferville (54.8°N, 66.8°W) has operated in a limited way since 1987. The Halley Station radar (76°S, 27°W) has essentially operated like the Goose Bay radar (except for the vertical interferometer capability) since January, 1988.

We now turn to consideration of the information that can be gleaned from the HF coherent scatter returns. For more detailed discussions, see *Villain et al.* [1987] and *Baker et al.* [1988].

Signal Processing

The on-line analysis yields autocorrelation functions (ACFs) with real (R) and imaginary (I) parts. Ideally, these are damped cosine and sine functions, respectively. It is possible to Fourier transform the ACFs into 32-point Doppler spectra. However, it is advantageous to estimate the mean Doppler velocity and spectral width through direct analysis of the ACF. This is done by performing a linear least-squares fit of the phase angle [arctan(I/R)] as a function of the lag number. Figure 5 presents an example of an ACF (5*a*) and its phase angle variation (5c). It can be seen that there is little deviation from the linear least-squares fit, resulting in an extremely small error on the estimate of the mean line-of-sight Doppler velocity. The single exception is lag 13 which is not processed for the reasons described by *Greenwald et al.* [1985].

The width of the Doppler spectrum can be inferred from the variation of the power $(R^2 + I^2)^{1/2}$ of the ACF as a function of lag number. We have found that the decorrelation of the ACF with time is generally best described by an exponential expression of the form

$$P(t) \quad = \quad C\, e^{i\omega t}\, e^{-\lambda t} \qquad [1]$$

where C is the zero lag power, ω the Doppler frequency, and λ the decorrelation parameter. The Fourier transform of P gives a Doppler spectrum of the form

$$S(\omega') \quad = \quad C\,\frac{2\lambda}{\lambda^2 + (\omega - \omega')^2} \qquad [2]$$

which peaks at $\omega' = \omega$ and has a full width at half maximum of 2λ rad/s.

Figure 5 illustrates the signal processing for a sample of HF radar data. Figure 5*a* shows the real and imaginary parts of the ACF. Figure 5*b* presents the Doppler spectrum that is obtained from the Fourier transformation. The phase angle variation shown in Figure 5c is best fit by a velocity of -423 m/s with an estimated error of 14 m/s. Figure 5d shows the variation of power with phase angle. Two fits to the variation are shown. That obtained from the decorrelation expression of Equation 1 is much better than the second fit, which was obtained from an expression quadratic in λ. The value of 30 Hz for λ corresponds to a half-power spectral width of 130 m/s. The vertical line on the spectrum of Figure 5*b* corresponds to the mean Doppler velocity computed from the fit of Figure 5c, and the horizontal bar shows the half-power width as computed from the fit of Figure 5d.

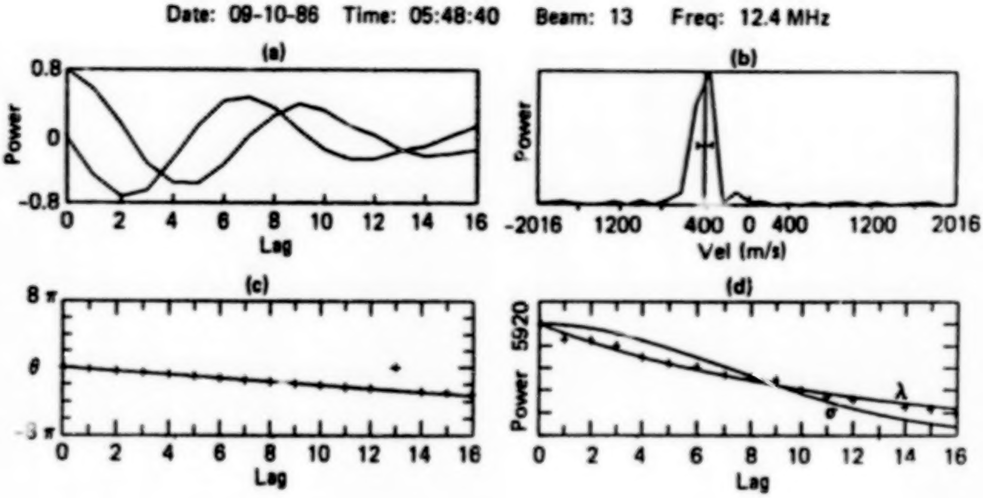


Figure 5. Stages in the processing of a sample of HF radar data. (*a*) Real and imaginary parts of an ACF. (b) Doppler spectrum obtained by Fourier transformation of the ACF. (c) Phase angle as a function of lag and its linear least-squares fit. (d) Power variation of the ACF as a function of lag and fits based on exponential λ and exponential $\sigma = \lambda^2$ as explained in the text. [*Reprinted from* Villain *et al*., 1987]

The Relation of Irregularity Drift To Plasma Drift

Generally, the signal backscattered to the HF radar is Doppler shifted. This corresponds to a relative line-of-sight motion of the phase fronts of the irregularity structures. Recently, studies in the high-latitude E region have demonstrated that the relationship between irregularity velocity and **E X B** velocity in that altitude regime is complex (i.e. *Nielsen and Schlegel* [1985]). However, studies with HF radars in the high-latitude F region indicate the direct equality of the irregularity velocity and the drift velocity of the ambient plasma [*Villain et al.*, 1985; *Ruohoniemi et al.*, 1987]. Figure 6 illustrates the results of a study carried out with the Goose Bay HF radar and the SRI incoherent scatter radar located at Sondrestrom, Greenland. Over a 40-minute period well-correlated reversals in the line-of-sight velocity of over 1 km/s were observed by the two radars.

It must be emphasized that the velocity measurement made by an HF radar in the F region is sensitive to only the plasma drift perpendicular to **B**, i.e. the **E X B** velocity. This is a consequence of the alignment of the irregularity phase fronts with the geomagnetic field lines. Thus, the motion of the irregularities in the high-latitude F region can be used to trace the ionospheric convective drift.

Localization of the HF Echo Sources

It has already been noted that ionospheric refraction is required for irregularities to be observed in the high-latitude F region. The exact path of the radar signal through the ionospheric medium is then unknown. It is important to determine how accurately the location of the echo sources can be fixed given the echo parameters at the radar, such as group delay, elevation angle, and azimuth angle. This problem has been addressed by *Villain et al.* [1984], who modelled HF ray propagation in a realistic high-latitude ionosphere derived from measurements made with an incoherent scatter radar. An example of a ray tracing is shown in Figure 7. The portions of the paths overprinted with heavy hatching indicate where the rays come to within 1° of orthogonality with **B**, i.e. where backscattering from small-scale irregularities is possible. Rays with low take-off angles (0-14°) reach orthogonality in the E region and are in fact entirely reflected there. Rays in the 20-30° interval penetrate the E and F regions without approaching orthogonality. In the 16-18° interval the rays penetrate the E region but come close to orthogonality in the F region. It is found that the error in the localization of the echo sources obtained by assuming a range to the scattering volume equal to the group delay range is generally less than about 50 km. This corresponds to only one range gate of uncertainty on the HF radar measurements. The sensitive calibration experiment described by *Ruohoniemi et al.* [1987] produced an empirical offset factor of just 15 km for echoes obtained from ranges of 1000 km.

IONOSPHERIC DYNAMICS

Mapping High-Latitude Plasma Convection

The HF radar technique is readily suited to studies of high-latitude plasma convection. In this section, we outline the methods developed to synthesize HF radar data into two-dimensional maps of convection velocity. A more detailed discussion can be found in *Ruohoniemi et al.* [1989].

Figure 8 shows examples of maps of velocity. Figure 8a presents line-of-sight velocity data for a 16-beam scan carried out by the Goose Bay radar. Also plotted are line-of-sight vectors for simultaneous Schefferville data. The Schefferville radar at this time (October, 1986) was operated on a single look direction. The limited region of overlap of line-of-sight

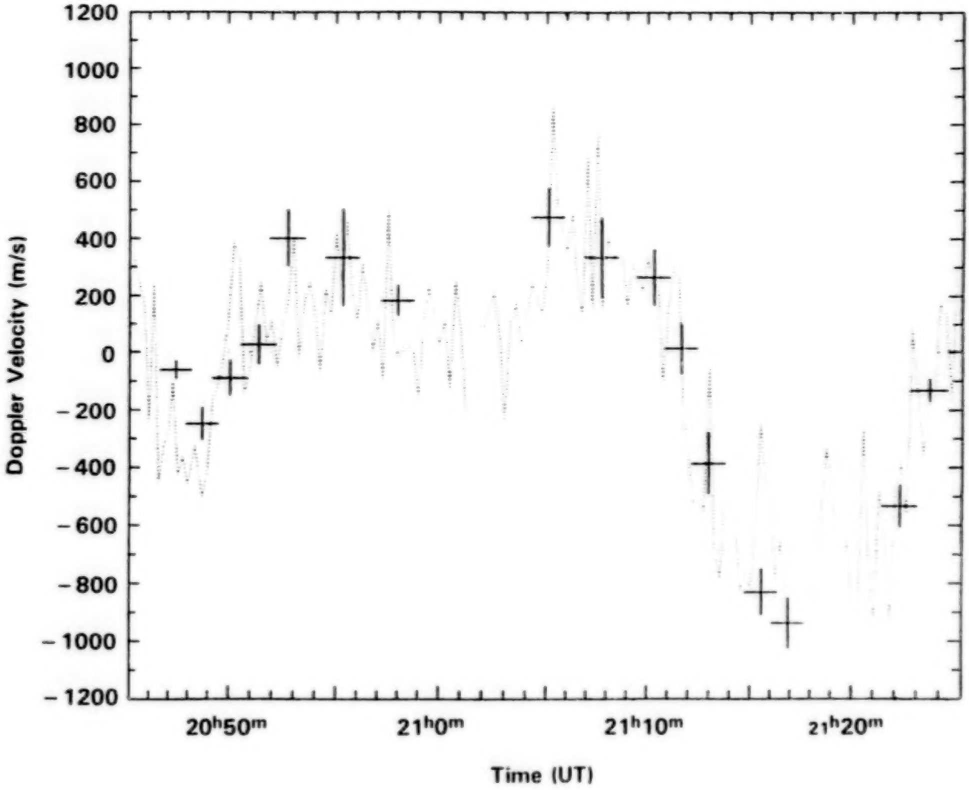


Figure 6. Comparison of time series of line-of-sight velocity estimates for a 40-minute period of coordinated Goose Bay - Sondrestrom observations. The 15-s incoherent scatter velocity data are plotted as a semi-continuous trace without error bars. [*Reprinted from Ruohoniemi et al.*, 1987]

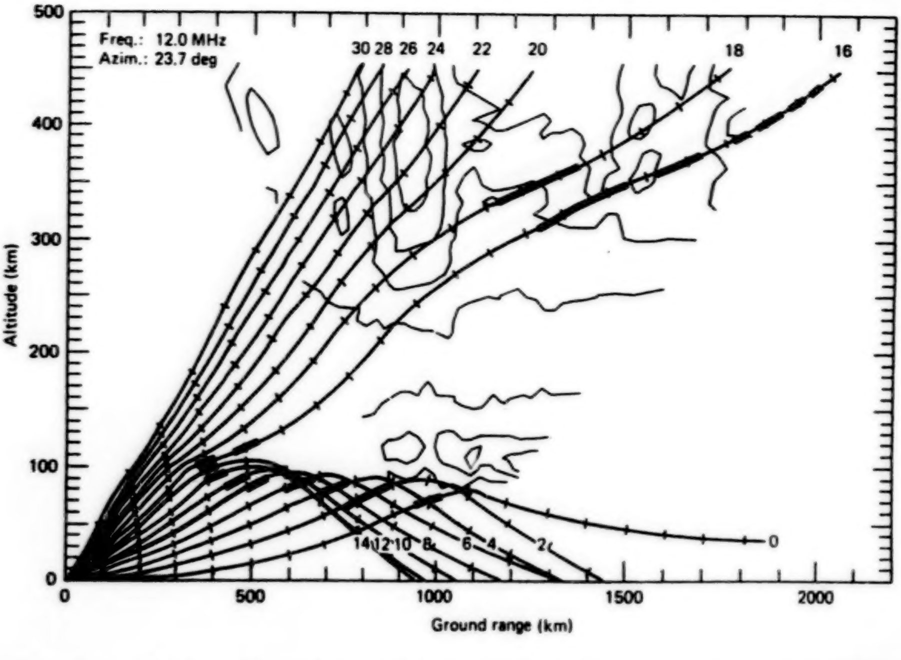


Figure 7. Ray tracing through a model ionosphere for a transmitted frequency of 12 MHz. Representative contours of the density distribution are shown. The tic marks along the traces indicate increments of 100 km in group delay. The heavy hatching shows where the rays come to within 1° of orthogonality to B. [*Reprinted from Villain et al.*, 1984]

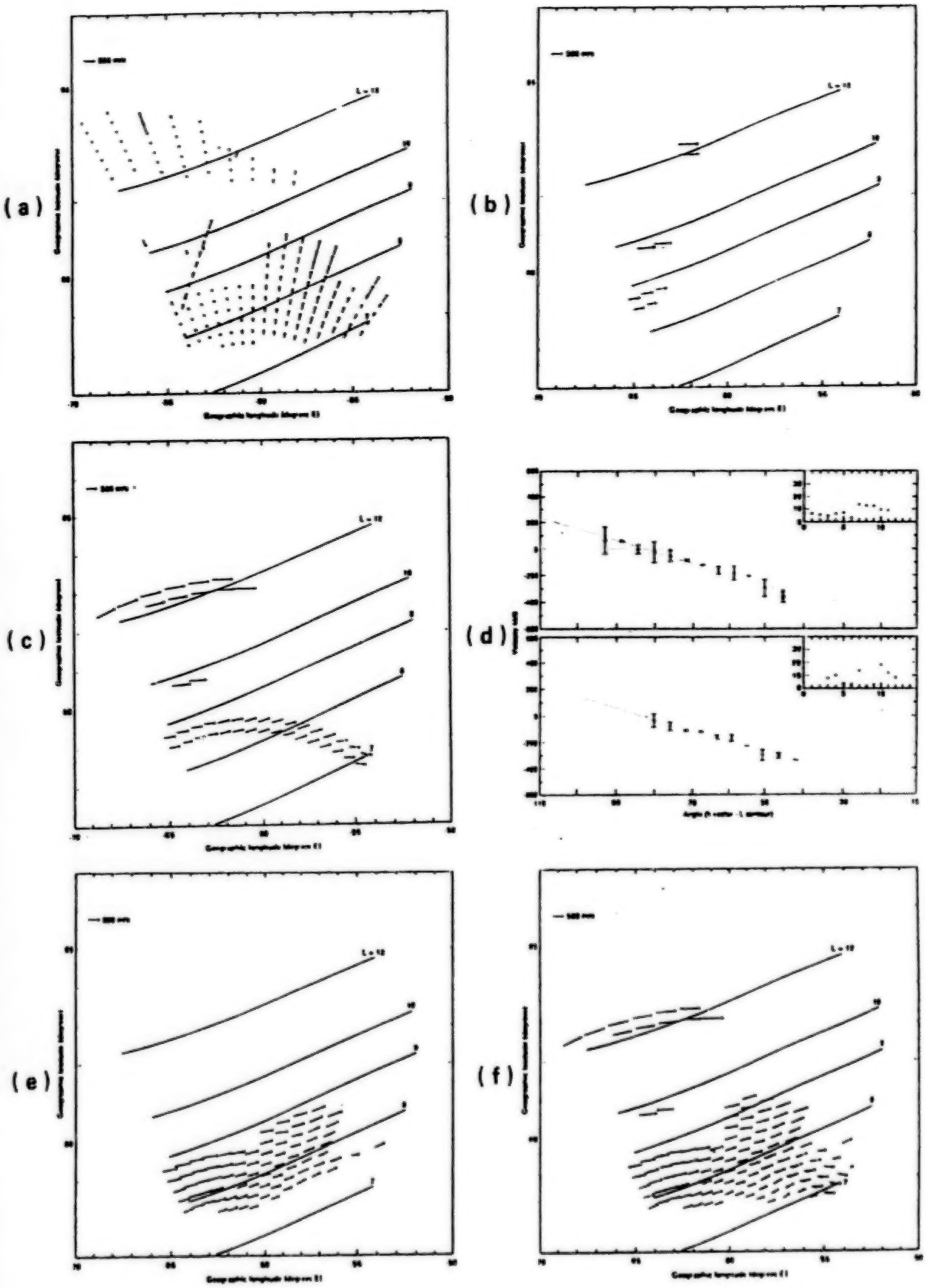


Figure 8. Sequence of plots illustrating the methods of mapping ionospheric plasma convection velocity using HF radars. For explanation, see text. [*Reprinted from Ruohoniemi et al.*, 1989]

velocity estimates can be solved vectorially for estimates of the convective drift velocity. The result is shown in Figure 8b. It is possible to extend this solution by noting that the overall velocity field must be divergence-free, i.e. $\nabla \cdot \mathbf{v} = 0$. The result is shown in Figure 8c. Finally, the variation of the line-of-sight velocity with azimuth may indicate that the plasma drift was constant in magnitude and direction along contours of constant L shell. Figure 8d shows the result of testing this condition on the Goose Bay data of this scan; excellent least-squares fits are obtained, indicating that the drift of plasma was almost due westward along L contours and of magnitude 500 m/s. Figure 8*e* shows the velocity map obtained by L shell analysis of the line-of-sight velocity data from Goose Bay alone. Figure 8*f* superposes the vectors of Figure 8c on those of Figure 8e extended by application of the divergence-free condition. It can be seen in the regions where the vectors are double that excellent agreement is obtained between the different methods of solution.

Thus, there are several techniques for analyzing data collected with HF coherent radars for information on ionospheric convection. These consist of (i) merging line-of-sight velocity data from widely separated radars, (ii) analyzing line-of-sight velocity data for uniformity of the motion with respect to contours of constant *L* shell, and (iii) applying the condition that the overall flow be divergence-free ($\nabla \cdot \mathbf{v} = 0$) to line-of-sight velocity data, with seed vector velocity data provided by method i) or ii).

Goose Bay - Sondrestrom Observations of the Polar Cleft

We illustrate the velocity mapping techniques with an example taken from a coordinated study of the polar cleft carried out with the Goose Bay radar and the SRI incoherent scatter radar located at Sondre Stromfjord, Greenland (67°N, 51°W). The incoherent scatter measurements provide estimates of the line-of-sight component of the bulk plasma motion and thus the SRI radar can serve as a second radar for vectorial analysis of the convection pattern. Figure 9*a* shows the overlapping maps of line-of-sight velocity estimates that were obtained during a six-minute period. It indicates that in addition to strong ($>$ 500 m/s) flows, both radars observed dramatic reversals in the convection. The result of directly merging the two sets of velocity estimates for velocity vectors is shown in Figure 9*b*. (The solution is limited to pairs of vectors that intersect at angles greater than 40°.) The principal motion is a northeasterly drift. With decreasing latitude through 66° the motion turns abruptly westward. Towards the highest latitudes ($>$ 70°) the drift rotates into a more poleward orientation.

The region of velocity vectors can be enlarged by applying the divergence-free ($\nabla \cdot \mathbf{v} = 0$) condition to both sets of line-of-sight velocity data. The vector solutions illustrated in Figure 9*b* provide vector velocity estimates for initializing this analysis. Figure 9c shows the result. It is clear that the region of northeasterly flow extends to earlier local times. At the highest latitudes the flow turns poleward and westward.

In this complicated system of flows, we discern two large-scale cells of convection. The poleward cell extends from earlier local times in the cleft period and includes a substantial flow towards the polar cap. The equatorward cell intrudes into the cleft period from the afternoon sector, where the flow is predominantly westward. The combination of a shear reversal at lower latitudes with a poleward turning of the flow at higher latitudes is consistent with the post-noon Heppner-Maynard convection patterns that are obtained for the northern hemisphere under conditions of southward IMF B_z and negative B_y [*Heppner and Maynard*, 1987]. Solar wind data for this period indicate that indeed these were the orientations of the IMF components.

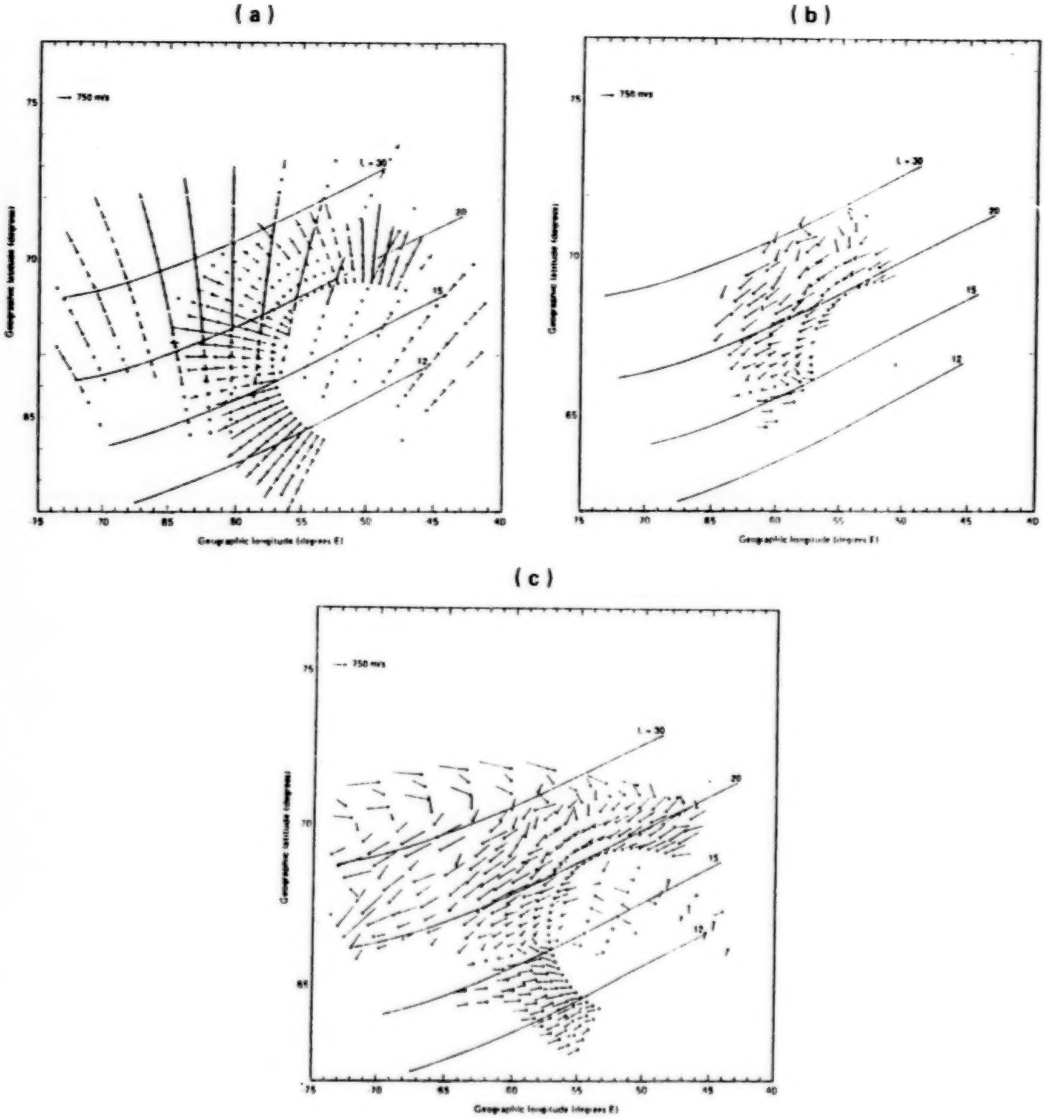


Figure 9. Sequence of plots showing the mapping of plama convection using HF coherent scatter data from Goose Bay and incoherent scatter data from the SRI radar at Sondre Stromfjord. For explanation, see text. [*Reprinted from Ruohoniemi et al.*, 1989]

IONOSPHERIC IRREGULARITIES

E Region Instability Studies: Electrostatic Ion Cyclotron Waves

The small-scale (~ 10 m) irregularities observed by coherent scatter HF radars are generated by plasma instabilities. In the high-latitude E region, the Farley-Buneman and gradient-drift instabilities are thought to account for the bulk of the observations (*see Fejer and Kelley* [1980]). However, a new type of echo has recently been identified that seems to relate to the electrostatic ion cyclotron (EIC) instability [*Villain et al.*, 1987]. We review the experimental evidence for the existence of these echoes and their origin in magnetic field-aligned electron-ion drifts.

An EIC-type event is characterized by the appearance of an L shell-aligned arc of echoes in the early morning period. The motions present within the arc tend to be slow and smoothly varying with azimuth, of such a sense as to indicate eastward motion. The most important feature is the appearance of localized regions of high Doppler velocity within the arc. The high Doppler velocities take on well-defined values in the range 350-650 m/s. Figure 10*a* presents a plot of the variation of the line of sight Doppler velocity with angle between *k* vector and L contour for a representative EIC event. The low velocity component is well-fitted by a cosine curve obtained for a drift almost due eastward along the L contour of magnitude 217 m/s. The high-velocity signal is superimposed on this smooth variation over two beams and reaches values slightly exceeding 500 m/s. The localized regions of high velocity last for periods of a few minutes to a few tens of minutes. Figure 10*b* shows examples of 'jumps' in line of sight Doppler velocity data plotted as a function of range for selected beams. From analysis of this sort, the existence of two types of high velocity signal can be inferred. One is distributed between 320 and 550 m/s and has an average value of 445 m/s, while the other is distributed between 500 and 650 m/s and has an average value of 580 m/s.

As discussed by *Villain et al.* [1987], the lower set of high Doppler velocities represents the ion acoustic velocity in the medium. Echoes with this phase velocity are associated with the Farley-Buneman type of instability. It is useful to plot the ratio of the two higher velocity components where their simultaneous presence could be identified from velocity jumps in the data base. This plot is shown in Figure 11. Also plotted are two curves representing the ratio of the Doppler velocities of EIC waves (V_{eic}) and the ion acoustic velocity, C_s. V_{eic} is computed from fluid theory and is controlled by three parameters: the ion cyclotron frequency Ω_i (which varies with ion mass), the irregularity wave vector *k*, and C_s. The ratio $R = V_{eic}/C_s$ is expressed analytically by

$$R = \frac{1}{C_s}\left[\frac{\Omega_i^2}{k^2} + C_s^2\right]^{1/2} \qquad [3]$$

The theoretical curves are drawn for two ion species, O^+ and NO^+. The O_2^+ curve would fall just below the NO^+ curve and the N_2^+ one just above. The excellent agreement between the experimental data and the theoretical curve for NO^+ indicates that the lower of the high-velocity modes is associated with the ion acoustic instability and the higher velocity component is associated with NO^+ EIC waves. The low velocity component (~ 200 m/s), from which the plasma drift can be inferred, appears to be related to the action of the gradient-drift instability.

It is significant that the convective plasma drift in these events is insufficient to excite the Farley-Buneman instability. *Villain et al.* [1989]

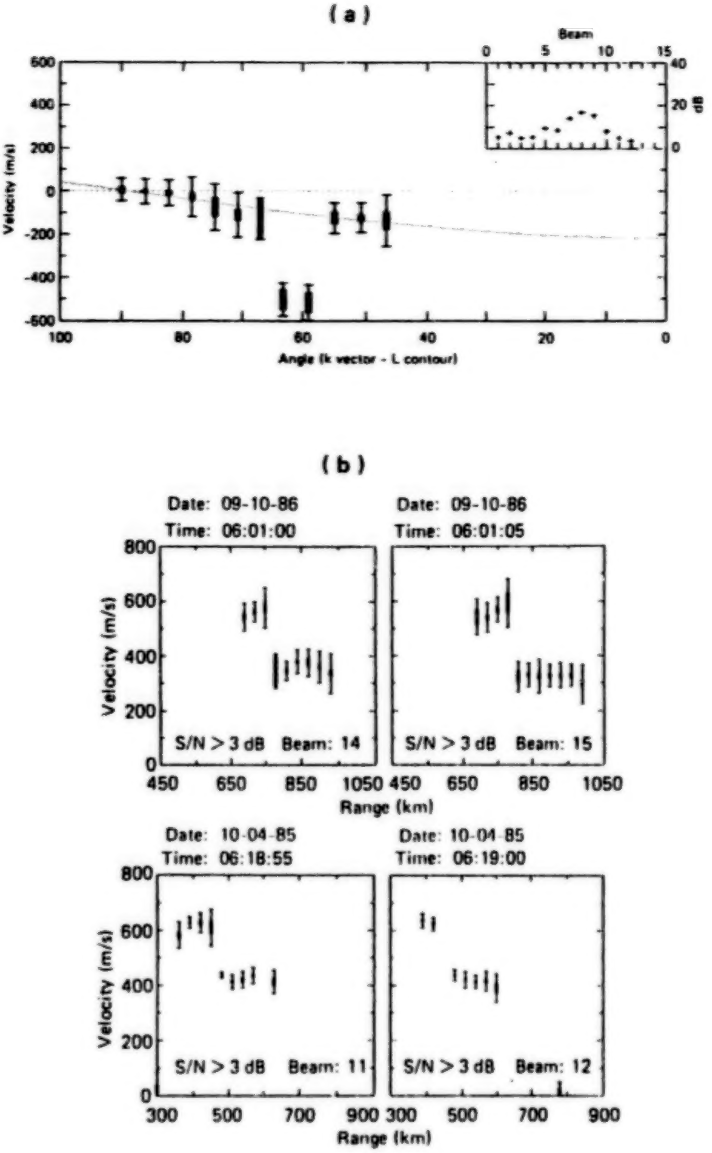


Figure 10. (*a*) Doppler velocity variation as a function of the angle between radar *k* vector and L contour for an EIC event. Note the presence of high Doppler velocity signals superposed on a background of low velocity signals. (b) Examples of transitions between high and low Doppler velocity signals. The plots show velocity as a function of range for selected beams. [*Reprinted from Villain et al.*, 1989]

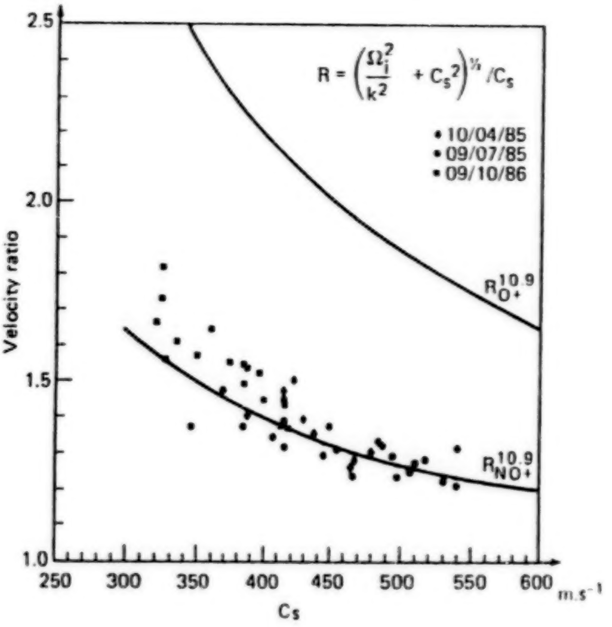


Figure 11. Ratio between the velocities observed on either side of velocity transitions plotted as a function of the lower velocity, assumed to be C_S. The theoretical value of this ratio is plotted for O^+ and NO^+ ions. [*Reprinted from Villain et al.*, 1989]

have presented direct evidence that ion acoustic modes can be produced in regions of subcritical perpendicular drift. The clear implication is that additional destabilization of the Farley-Buneman modes occurs as a result of the field-aligned drifts. Thus, field-aligned currents may play an important role in the interpretation of 'classical' Farley-Buneman echoes as well as the more recently discovered EIC echoes.

F Region Instability Studies: The Wintertime Trough at Dusk

It has been found that there are strong diurnal and seasonal dependences in the occurrence of coherent scattering from the high-latitude ionosphere. In this section, we describe a type of activity seen with remarkable regularity in the dusk sector of the wintertime subauroral ionosphere. For a more detailed discussion of this phenomenon see *Ruohoniemi et al.* [1988].

Figure 12 shows a sequence of images obtained with the Goose Bay radar during a one-hour period of observations in the late afternoon/early evening of January 12, 1986. The feature of interest is the band of echoes just off the coast of the Quebec-Labrador land mass. It first appeared (1949:25 UT) as a few weak echoes near the eastern edge of the field of view. With time it intensified and moved westward (2011:00 UT). It was possible to track the westward movement of the most intense part of the echo activity (2021:35 UT). Eventually, the scatter weakened in the eastern sector as the activity was observed to rotate entirely out of the field of view (2037:25 UT).

This example is illustrative of a class of scatter activity called dusk scatter that is observed by the Goose Bay radar with great regularity in the wintertime ionosphere, especially during quiet geomagnetic conditions. Dusk scatter first appears when the solar zenith angle in the scattering region is near 95°. It appears to move in a westerly direction and is observed for 1 - 1½ hours. The occurrence of dusk scatter is strongly dependent on season. Figure 13*a* shows a histogram of the monthly probability of dusk scatter occurrence. For the winter months the rate of occurrence is virtually 100%, while for the summer months such activity is almost absent. Figure 13*b* shows another manifestation of the influence of season on dusk scatter occurrence. Here, the average period of occurrence for 25-day intervals is plotted as a function of day number relative to the day of winter solstice. Also plotted are contours of constant solar zenith angle ($z = 95°, 105°$) for a point lying in the latitude interval where dusk scatter is observed. It can be seen that these curves essentially constitute bounds for the occurrence of dusk scatter. Thus, there is a strong solar zenith angle factor in the time of occurrence of dusk scatter.

As elaborated in *Ruohoniemi et al.* [1988], the irregularities responsible for dusk scatter lie in the subauroral ionosphere in the range of latitudes associated with the trough. The convective plasma drifts in that region in geomagnetically quiet times tend to be low ($\lesssim$ 150 m/s) and westward in the sun-earth inertial reference frame. Density gradients associated with the trough at the afternoon solar terminator are favorable for the generation of irregularities through the gradient-drift instability. Thus, the trough at dusk is identified as a source region for small-scale irregularities.

GLOBAL MAGNETOSPHERIC DYNAMICS

Conjugate Observations of Dayside Convection

With the techniques developed to map convection with HF coherent scatter radars, it is possible to explore many questions pertaining to the nature of the magnetospheric response to the solar wind. The installation of the HF facility at Halley Station has added a dimension in global coverage [*Baker et al.*, 1989]. In this section, we briefly discuss the results of a

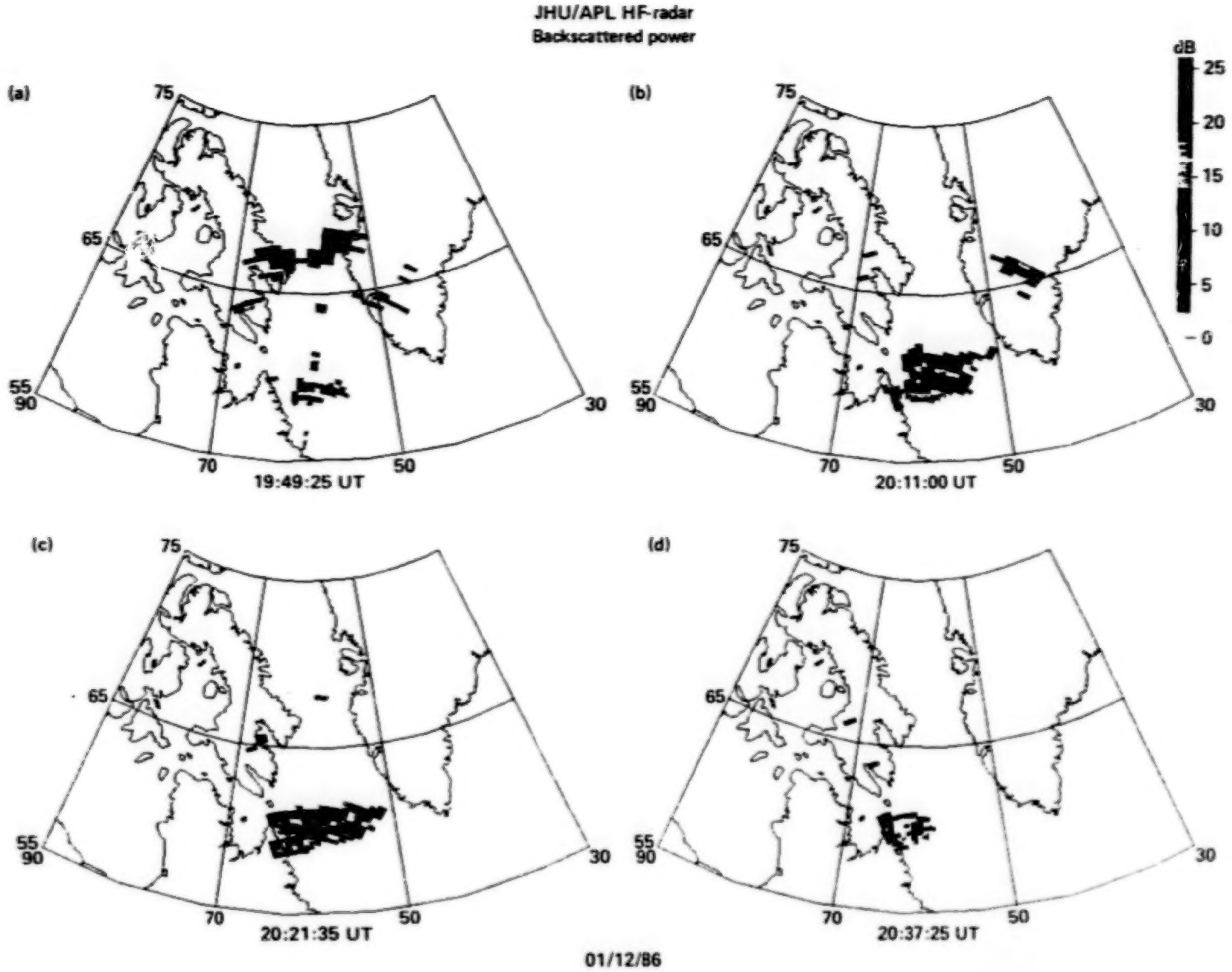


Figure 12. Set of four maps of backscattered power showing the development of HF backscattering activity near dusk on January 12, 1986. [*Reprinted from Ruohoniemi et al.*, 1988]

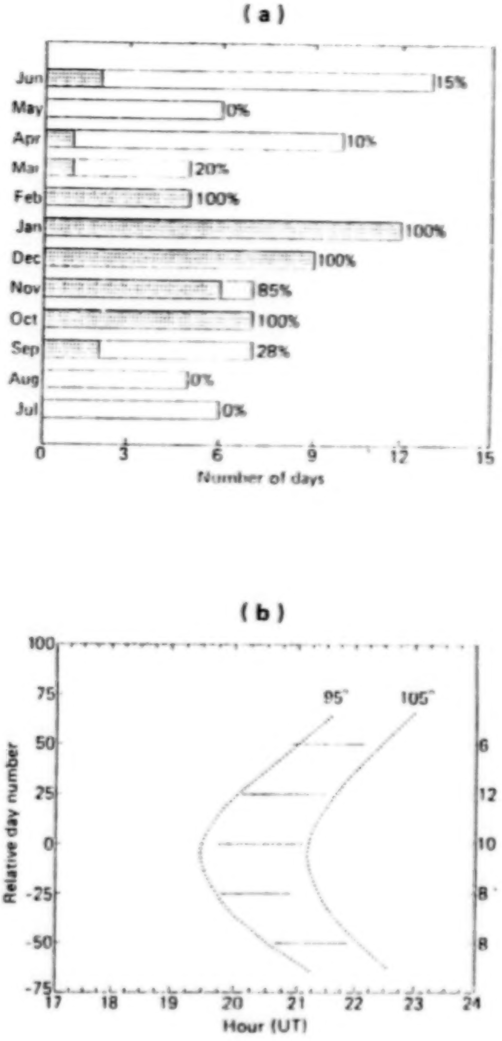


Figure 13. (*a*) Monthly occurrence of dusk scattering events for magnetically quiet conditions ($K_p \lesssim 2$). The outer bar indicates the number of quiet days in each month. The inner shaded bar shows the number for which dusk scatter was recorded. The percentage indicates the monthly occurrence probability of dusk scatter. (*b*) The average period of dusk scatter for 25-day intervals centered on the day of winter solstice. The numbers along the right-hand margin indicate the number of events contributing to the calculation of each average. The dotted curves are contours of constant solar zenith angle ($z = 95°$, $105°$) for a point lying due north of Goose Bay at the range of dusk scatter. [*Reprinted from Ruohoniemi et al.*, 1988]

coordinated Goose Bay - Halley Station study of the response of dayside convection to the IMF. A more detailed exposition may be found in *Greenwald et al.* [1989].

First, we note that the Halley Station and Goose Bay radars have a large common conjugate viewing area. Figure 14 shows this common area with the location of Goose Bay and its field of view projected onto the southern hemisphere. The conjugate mapping is performed with the aid of a magnetic coordinate system especially developed for conjugate studies by *Baker and Wing* [1989]. The conjugate area is close in size to the individual field of view of either radar ($\sim 3\ 10^6\ km^2$).

The observations described here were obtained during a geomagnetically disturbed period in which the radars were located in the noon local time sector. Magnetic field measurements were made in the solar wind by the IMP-8 satellite; plots of the IMF data are presented in Figure 15. During the period of interest (1445-1600 UT) B_y underwent a number of reversals while B_z remained negative. It is interesting to consider the convection patterns predicted by the Heppner-Maynard model for these conditions (see *Heppner and Maynard* [1987]). These are shown in Figure 16 for the northern hemisphere. Note that the convection patterns predicted for the southern hemisphere are complementary in the sense that the northern pattern for $B_y<0$ applies to the southern hemisphere for $B_y>0$ and vice versa. The dominant feature of the the northern hemisphere $+B_y$ pattern is the strong westward flow across local noon into a throatlike region in the prenoon sector. The principle characteristic of the $-B_y$ pattern is the clockwise rotation of the flow in the post noon sector with increasing latitude.

We now turn to maps of the convection velocity obtained with the HF radars. These are shown in Figure 17. In Figure 17*a* the Goose Bay data show uniform poleward and westward flow across the entire local noon time sector. The Halley Station map shows a sharp clockwise turning of the flow with latitude. The IMF data indicate that the B_y component was positive during this time interval (1530-1540 UT). The match with the Heppner-Maynard patterns for this orientation of B_y is very close. Figure 17*b* shows the convection maps obtained for an interval in which B_y was negative (1453-1511 UT). It is clear that the convection patterns are exchanged between the hemispheres on change of sign of the B_y component, as predicted by the Heppner-Maynard model.

We conclude that in this event the instantaneous convection patterns observed in conjugate hemispheres were consistent with the statistical patterns of *Heppner and Maynard* [1987]. In addition, we can utilize the fine temporal resolution (~ 1½ min) of the HF radar observations to study the details of the transition between convection states. As discussed in *Greenwald et al.* [1989] it is found that the average time between an IMF transition at the IMP-8 satellite and the beginning of a convection reconfiguration in the ionosphere was 8 minutes. The newly reconfigured convection pattern typically filled the radar field of view (10°-15° of invariant latitude and 2½ hours of local time) within 6 minutes of reconfiguration onset.

The studies with Halley Station are continuing and promise to touch on very basic questions regarding the global dynamics of the magnetosphere and hemispheric conjugacy.

IONOSPHERE-THERMOSPHERE COUPLING

Goose Bay Observations of Atmospheric Gravity waves

The previous examples involved measurements of the drift of small-scale irregularities within the convecting ionosphere. We now turn to a study in which the motion of ionospheric layers themselves is directly measured by the

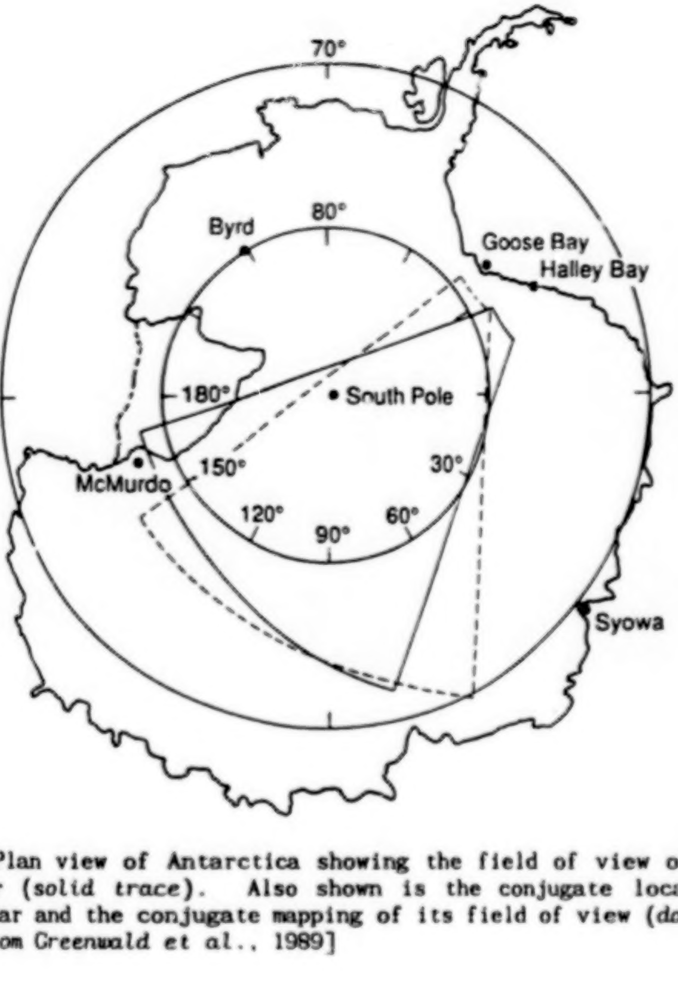


Figure 14. Plan view of Antarctica showing the field of view of the Halley Station radar (*solid trace*). Also shown is the conjugate location of the Goose Bay radar and the conjugate mapping of its field of view (*dashed trace*). [*Reprinted from Greenwald et al.*, 1989]

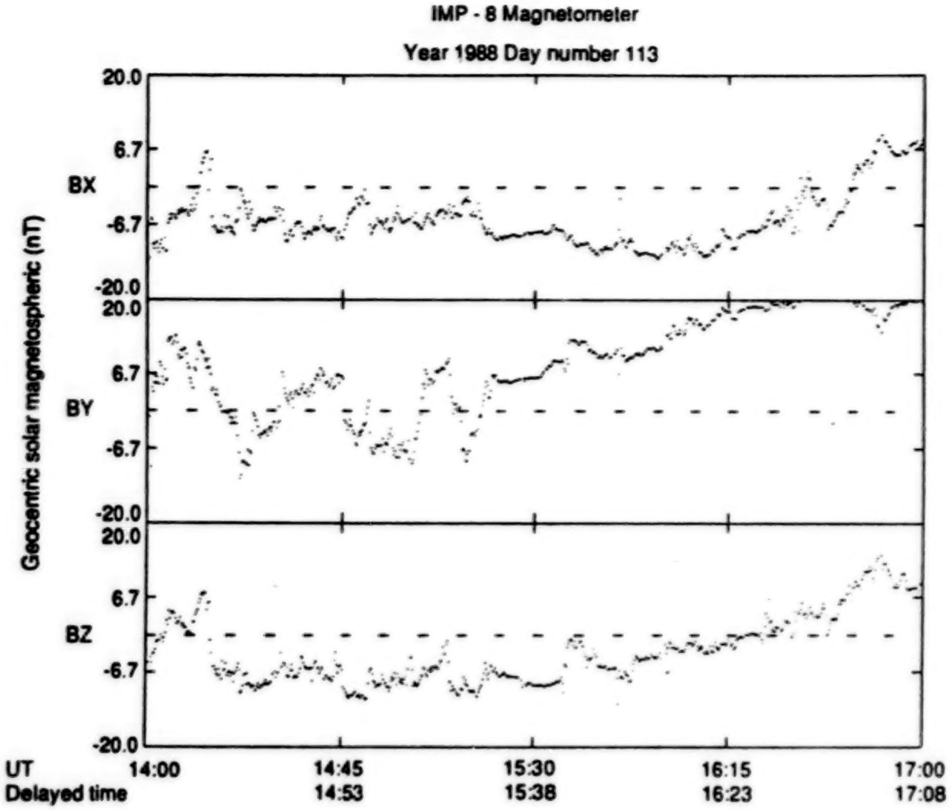


Figure 15. Magnetometer data from the IMP-8 satellite for April 22, 1988, 1400-1700 UT. A second "delayed time" axis is obtained by adding 8 minutes to the UT time. This delay is the empirically determined response time between an IMF B_y state change at the satellite and the onset of a convection response in the ionosphere. [*Reprinted from Greenwald et al.*, 1989]

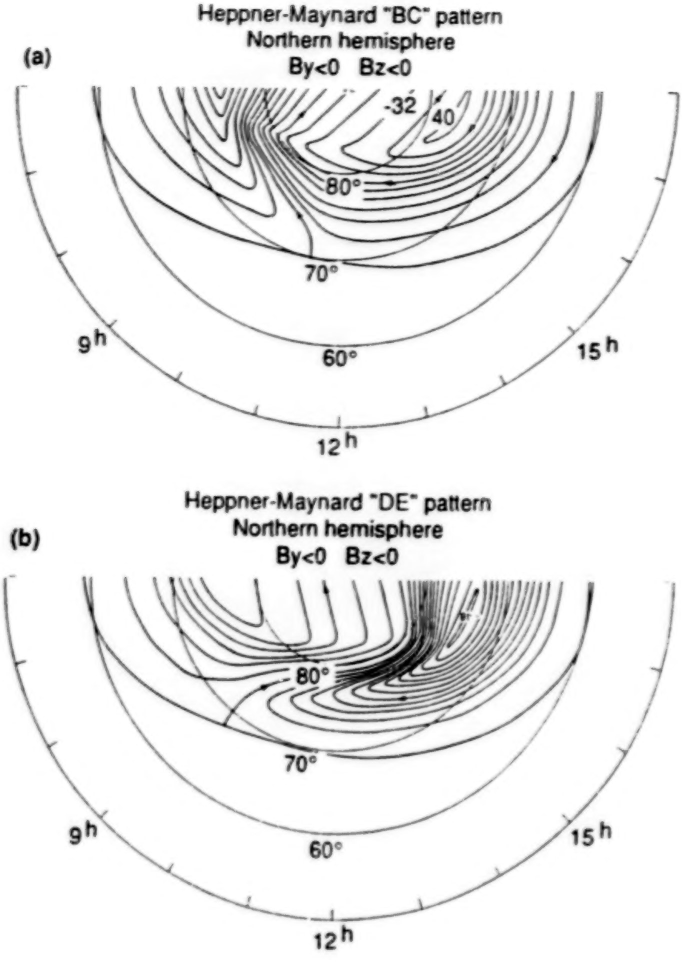


Figure 16. Dayside high-latitude ionospheric convection patterns derived by *Heppner and Maynard* [1987] for IMF $B_z<0$ conditions: (*a*) BC pattern for northern hemisphere with IMF $B_y>0$ or southern hemisphere with IMF $B_y<0$. (*b*) DE pattern for northern hemisphere with IMF $B_y<0$ or southern hemisphere with IMF $B_y>0$. [*Reprinted from Greenwald et al.*, 1989]

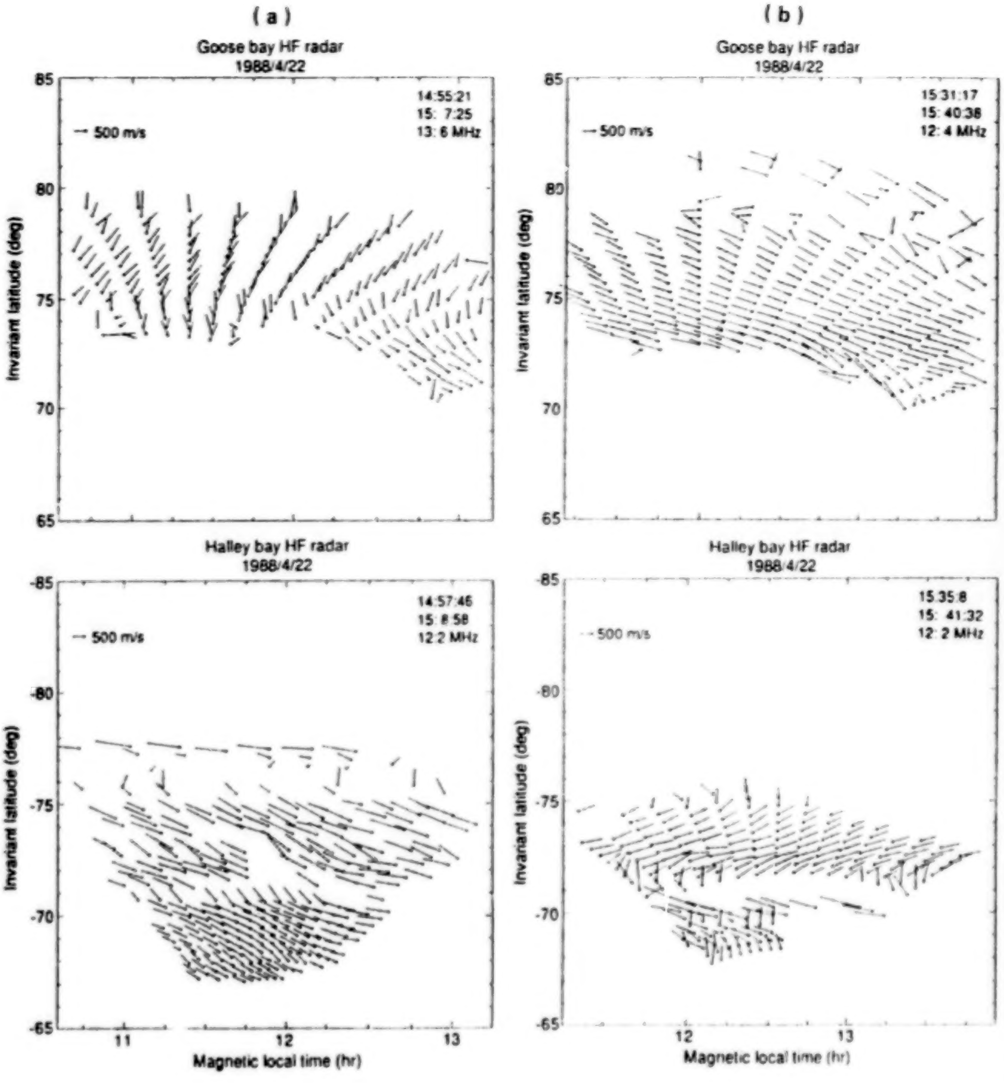


Figure 17. Simultaneous convection patterns in the dayside ionosphere as observed with the Goose Bay and Halley Station radars. (*a*) Period for which $B_y>0$. (b) Period for which $B_y<0$. [*Reprinted from Greenwald et al.*, 1989]

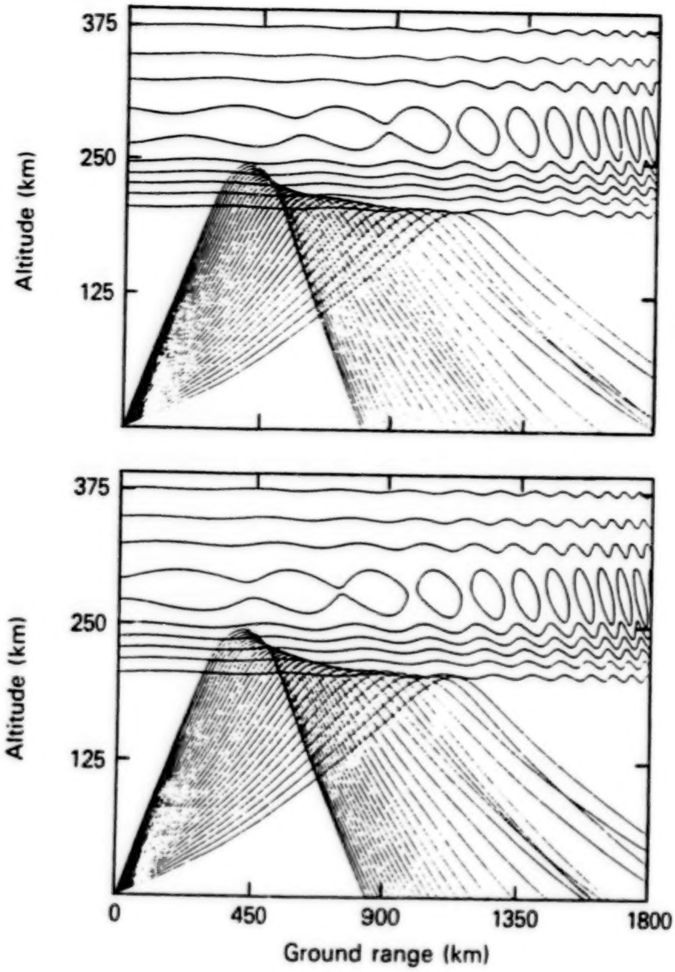


Figure 18. Ray paths of HF signal through an ionosphere modulated by an earth-reflected gravity wave. Note the focusing caused by the surfaces of the density minima. Top: 195 minutes after excitation; Bottom: 210 minutes after excitation. [*Reprinted from Samson et al.*, 1989b]

Goose Bay radar. Furthermore, we will show that the motion of the layers is due to the passage of gravity waves in the neutral atmosphere and derive important parameters describing the propagation of the gravity waves. This work has been described in detail by *Samson et al.* [1989*a*] and *Samson et al.* [1989*b*].

Gravity waves propagating in the neutral ionosphere induce electron density perturbations that can be likened to corrugation-type fluctuations in the ionospheric layers. Figure 18 illustrates the propagation of HF radar signals in an ionosphere disturbed by the passage of a gravity wave. The ray paths in Figure 18*a* show three regions where the rays are focused. The first is the usual focus near the mimimum 'skip' distance, while the second and third are due to focusing of the rays that are reflected by the corrugation minima in the perturbed electron densities. When the radar waves are backscattered by the ground, the focused rays show up as enhanced intensities at the corresponding ranges. Note that the regions of localized intensification are related to specific phase surfaces of the gravity wave. Figure 18*b* illustrates the ray path focusing after a delay of 15 minutes. The focusing from the more equatorward electron density minimum is no longer obvious since it has merged with the skip distance focusing. The more poleward region is focused at a nearer range and a new region of focusing is forming at a more distant range.

Figure 19*a* shows the theoretical prediction for the movement of constant phase surfaces of the gravity wave from a high latitude source region toward the Goose Bay radar. This range-time plot applies to observations on one beam of the Goose Bay radar. From this theoretical phase map, it is possible to determine the time and range at which the gravity wave was generated by extending the measured phase surfaces to the position where they intersect on the range-time plane. The same procedure can be followed with experimental data. Once the direction of propagation, range, and time have been estimated, the altitude of the source can be determined by comparing the data with phase surfaces computed for sources at different altitudes.

We now consider a sample of Goose Bay data that has been subjected to the analysis outlined above. Figure 19*b* shows a filtered range-time plot of backscattered intensity obtained from observations made with the Goose Bay radar on November 25, 1988. Also plotted are the theoretical surfaces of constant phase for a gravity wave event scaled to have the time of onset (1211 UT) at the range (2100 km) indicated by the radar data and an altitude of 125 km. The fit is excellent. Figure 19*c* shows that the time and location of the onset coincided with the development of a strong velocity shear in the vicinity of the high-latitude convection reversal boundary. Thus, it would appear that a strong equatorward-propagating gravity wave was initiated by a shear reversal intensification in the high-latitude electrojet.

Many parameters describing the propagation of gravity waves can be derived. For example, it has been found that the waves propagate very nearly equatorward across contours of constant L shell from high-latitude source regions. Occasionally, waves can be seen propagating poleward from the convection reversal region. The frequencies associated with the waves cover the range 0.3-0.6 mHz; the wavelengths, 300-500 km, and the average phase velocities, 110-180 m/s. The waves propagate in a mode that penetrates through the lower atmosphere and then reflects from the earth's surface. The vertical motions of the layers that accompany passage of the gravity waves can be measured from the slight motions imparted to the signal by the movement of the reflecting layers. One interesting point is that the gravity waves tend to be seen by the Goose Bay radar in the late fall and early winter. The reason may be that conditions for localized enhancements of the high-latitude electrojet are more favorable when particle precipitation can act on an ionosphere in relative darkness.

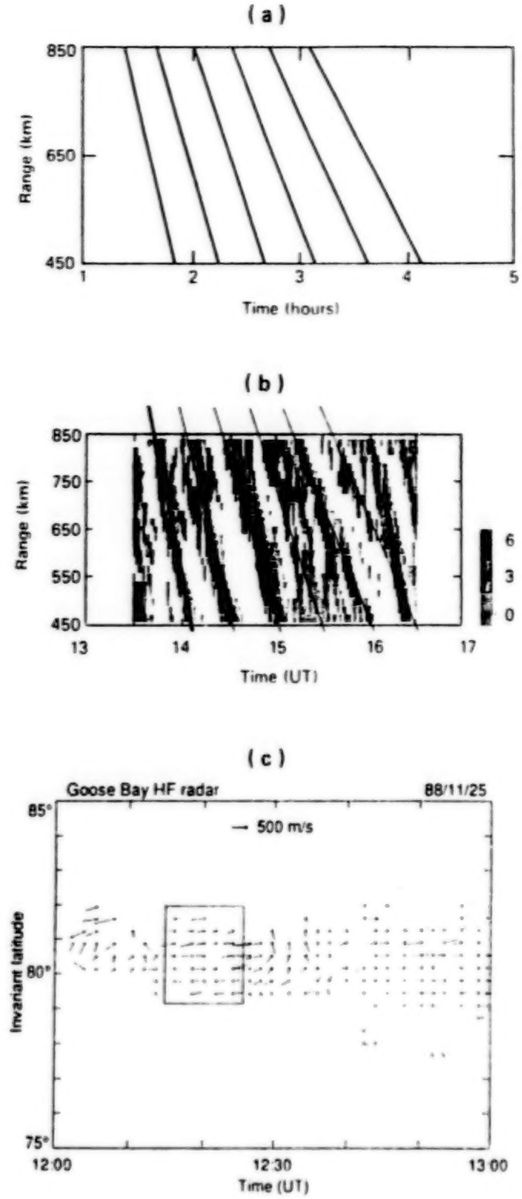


Figure 19. (*a*) Theoretical phase map for an earth-reflected gravity wave with a source at a range of 2000 km and an altitude of 130 km. The observation altitude is assumed to be 250 km. (*b*) A comparison of the best-fit theoretical phase map with the power recorded on a Goose Bay radar beam during a gravity wave event. (*c*) Time-series plot of the latitudinal profile of ionospheric convection for the period preceding the Goose Bay gravity wave event. The formation of the indicated velocity shear coincides with the onset of the gravity wave at very high latitudes. [*Reprinted from* Samson *et al.*, 1989b]

It is clear from the foregoing that HF coherent scatter radars are a powerful tool for the investigation of high-latitude gravity wave phenomena.

CLOSING COMMENT

The HF coherent scatter technique is broadly applicable to the goals of the WITS program, especially as concerns the study of ionospheric-magnetospheric dynamics, ionospheric plasma instabilities, and ionosphere-thermosphere coupling. The scope for this work has been considerably increased through the addition of an HF radar facility in Antarctica. The operation of these HF radars is extremely flexible and adaptable to the interests of researchers in the community. This flexibility extends to real-time control of the radar and monitoring of the backscattering activity during campaign periods. The results presented in this brief survey represent only a sampling of the scientific potential of the instruments. Many other topics remain for investigation. The authors welcome the involvement of WITS researchers in the planning, execution, and analysis of future HF radar experiments.

ACKNOWLEDGEMENTS

The Applied Physics Laboratory HF radar at Goose Bay, Labrador is supported in part by the National Science Foundation (NSF) Division of Atmospheric Sciences and the Air Force Office of Scientific Research, Directorate of Atmospheric and Chemical Sciences, under NSF grant ATM-8506851, in part by the National Aeronautics and Space Administration (NASA) under NASA grant 7055, and in part by the Defense Nuclear Agency and Rome Air Development Center under contract N00039-87-C-5301.

REFERENCES

Baker, K. B., R. A. Greenwald, J. P. Villain, and S. Wing, Spectral characteristics of high frequency (HF) backscatter for high latitude ionospheric irregularities: preliminary analysis of statistical properties, *Scientific report to the Rome Air Development Center, RADC-TR-87-284*, 1988.

Baker, K. B., R. A. Greenwald, J. M. Ruohoniemi, J. R. Dudeney, M. J. Pinnock, N. Mattin, and J. M. Leonard, PACE: Polar Anglo-American conjugate experiment, in press, *EOS*, 1989.

Baker, K. B., and S. Wing, A new magnetic system for conjugate studies at high latitudes, *J. Geophys. Res.*, *94*, 9139, 1989.

Bates, H. F., and P. R. Albcc, Aspect sensitivity of F-layer HF backscatter echoes, *J. Geophys. Res.*, *75*, 165, 1970.

Fejer, B. J. and M. C. Kelley, Ionospheric irregularities, *Rev. Geophys. Space Phys.*, *18*, 401, 1980.

Greenwald, R. A., K. B. Baker, R. A. Hutchins, and C. Hanuise, An HF phased-array radar for studying small-scale structure in the high-latitude ionosphere, *Radio Sci.*, *20*, 63, 1985.

Greenwald, R. A., K. B. Baker, J. M. Ruohoniemi, J. R. Dudeney, M. Pinnock, N. Mattin, J. M. Leonard, and R. P. Lepping, Simultaneous conjugate observations of dynamic variations in high-latitude dayside convection due to changes in IMF B_y, submitted, *J. Geophys. Res.*, 1989.

Heppner, J. P. and N. C. Maynard, Empirical high-latitude electric field models, *J. Geophys. Res.*, *92*, 4467, 1987.

Keskinen, M. J., and S. L. Ossakow, Theories of high-latitude ionospheric irregularities: A review, *Radio Sci*, *18*, 1077, 1983.

Nielsen, E., and K. Schlegel, Coherent radar doppler measurements and their relationship to the ionospheric electron drift velocity, *J. Geophys. Res.*, 90, 3498, 1985.

Ruohoniemi, J. M., R. A. Greenwald, K. B. Baker, J. P. Villain, and M.A. McCready, Drift motions of small-scale irregularities in the high latitude F region: an experimental comparison with plasma drift motions, *J. Geophys. Res.*, 92, 4553, 1987.

Ruohoniemi, J. M., R. A. Greenwald, J.-P. Villain, K. B. Baker, P. T. Newell, and C.-I. Meng, Coherent HF radar backscatter from small-scale irregularities in the dusk sector of the sub-auroral ionosphere, *J. Geophys. Res.*, 93, 12871, 1988.

Ruohoniemi, J. M., R. A. Greenwald, K. B. Baker, J.-P. Villain, C. Hanuise, and J. Kelly, Mapping high-latitude plasma convection with coherent HF radars, in press, *J. Geophys. Res.*, 1989.

Samson, J. C., R. A. Greenwald, J. M. Ruohoniemi, and K. B. Baker, HF radar observations of atmospheric gravity waves in the high-latitude ionosphere, in press, *Geophys. Res. Lett*, 1989*a*.

Samson, J. C., R. A. Greenwald, J. M. Ruohoniemi, A. Frey, and K. B. Baker, Goose Bay radar observations of earth-reflected atmospheric gravity waves in the high-latitude ionosphere, submitted, *J. Geophys. Res.*, 1989*b*.

Villain, J. P., R. A. Greenwald, and J. F. Vickrey, HF ray tracing at high latitudes using measured meridional electron density distributions, *Radio Sci.*, *19*, 359, 1984.

Villain, J. P., G. Caudal, and C. Hanuise, A SAFARI-EISCAT comparison between the velocity of F region small-scale irregularities and the ion drift, *J. Geophys. Res.*, 90, 8433, 1985.

Villain, J. P., R. A. Greenwald, K. B. Baker, and J. M. Ruohoniemi, HF radar observations of E region plasma irregularities produces by oblique plasma streaming, *J. Geophys. Res.*, 92, 12327, 1987.

Villain, J. P., C. Hanuise, R. A. Greenwald, K. B. Baker, and J. M. Ruohoniemi, Obliquely propagating ion acoustic waves in the auroral E-region; further evidence of irregularity production by field-aligned electron streaming, submitted, *J. Geophys. Res.*, 1989.

Chapter 13

THE DIGISONDE 256 IONOSPHERIC SOUNDER

Bodo W. Reinisch, Klaus Bibl, David F. Kitrosser, Gary S. Sales, Jane S. Tang,
Zhao-Ming Zhang, Terence W. Bullett and John A. Ralls

University of Lowell Center for Atmospheric Research, 450 Aiken Street, Lowell, MA 01854,
USA

1. INTRODUCTION

Some 60 years ago, the first ionosondes were developed based on the experiments by Breit and Tuve [1926]. These were simple instruments that measured the time of flight of pulse-modulated HF signals. In this paper we describe the new scientific applications that the modern digital ionosonde offers. We limit our discussion to the Digisonde 256, since it is the only globally used advanced ionosonde system. The DGS 256 network of some 40 sounders is shown in Figure 1. Nineteen of these sounders are being deployed by the U.S. Air Weather Service and the installation phase should be completed by the end of 1990 [Buchau, 1989]. The Digisonde 256 measures all observables of the ionospheric echoes [Reinisch, 1986]: amplitude, phase, range, frequency, Doppler shift, Doppler spread, angle of arrival, and wave polarization. Real time autoscaling of the ionogram data and calculation of the electron density profiles in ARTIST (Automatic Real Time Ionogram Scaler with True Height) eliminates some of the previously existing difficulties of ionosonde research: the tedious process of ionogram reduction and the time delay and costs associated with the calculation of the electron density profiles. Since the Digisonde can be remotely operated via standard telephone lines it is now relatively easy to prepare a station for special campaigns. A number of analysis techniques that may be useful for the Worldwide Ionosphere/Thermosphere Study have been developed over the last years and are described in Sections 3 to 7. As in the case of incoherent scatter radar research, it requires the involvement of many scientists to develop advanced analysis techniques and the corresponding software to make more extensive use of the observational techniques of the Digisonde 256. Section 2 summarizes the measuring techniques of the Digisonde system; for details we refer to the Digisonde Manual [Bibl et al., 1981].

2. THE DIGISONDE 256 SYSTEM

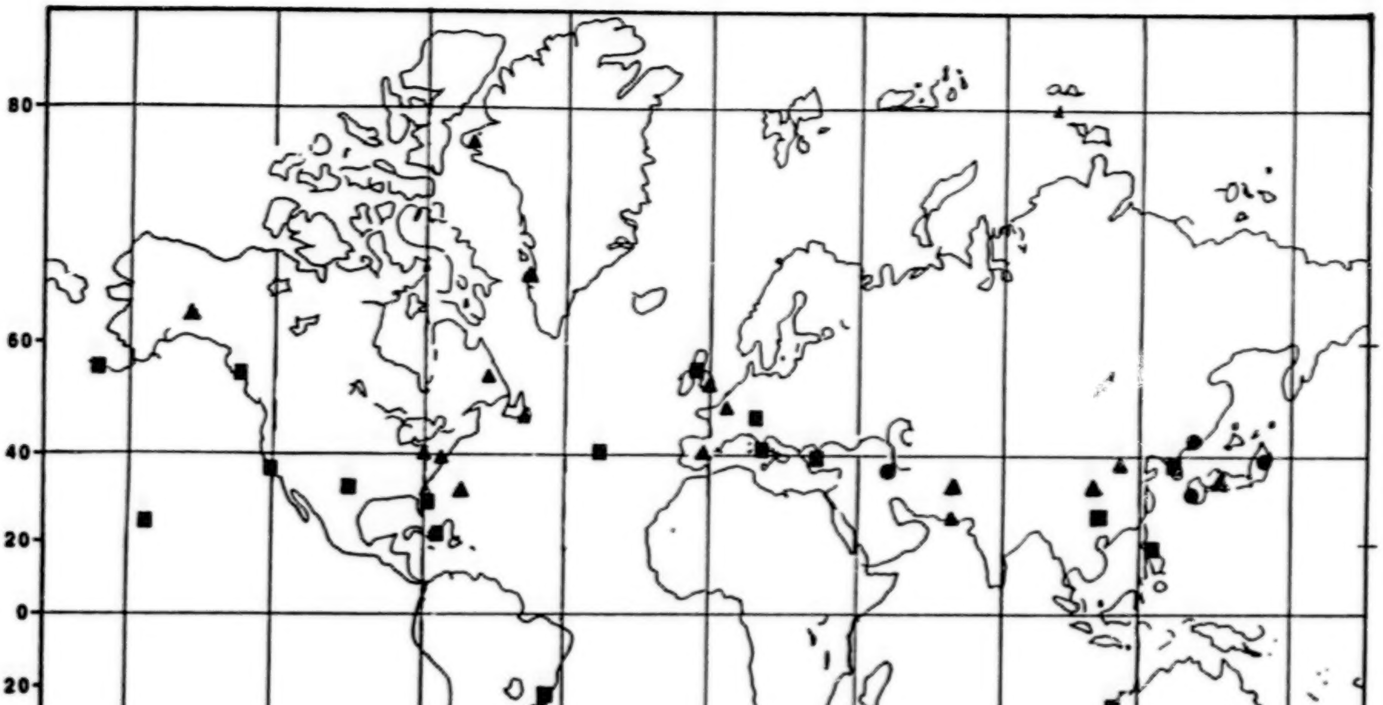
General Layout

Designed as a pulse radar system with a wideband 10 kW peak power transmitter and precise fast-switching frequency synthesis, the Digisonde 256 covers a frequency range from 0.5 to 30 MHz. A wideband transmit antenna with a vertical beam aperture of about 90° for vertical sounding and a log-periodic antenna for oblique and backscatter sounding are used in alternation. For reception an array of seven spaced crossed-loop antennas allows the polarization and the angle of arrival of ionospheric echoes to be measured. The array is laid out as an equilateral triangle with one antenna at its center; the aperture is nominally 100 m.

The sounder is installed in a rack and an equipment desk (Figure 2). The six foot rack houses the 10 kW transmitter, the uninterruptible power supply, and the antenna switch to which the seven receiving antennas are connected. The desk holds the transceiver/processor, ARTIST, tape drive, printer, and modems.

System Operation

The system operates in a number of different modes: ionogram, fixed frequency, and drift. In the ionogram mode, the sounder steps through the selected frequency band in increments of 5, 10, 25,



Fi[illegible]k of Digiso[illegible] 256 digital ionospheric sounders as of September

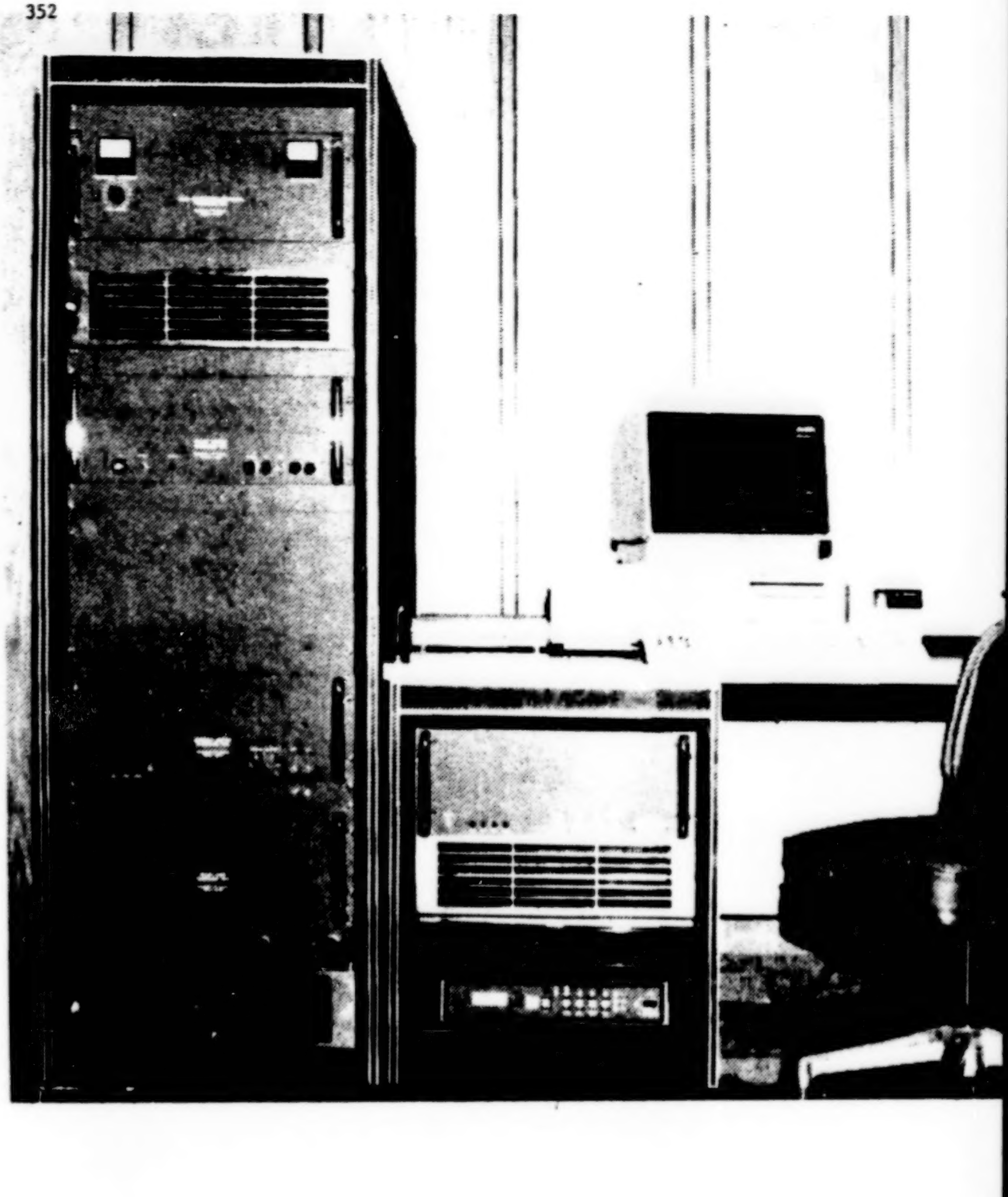


Figure 2 -- The Digisonde 256. The six foot tall rack on the left contains the 10 KW final amplifier and its pulsed power supply, the 7 receive antenna switch and the uninterruptible power supply. The desk rack (center) contains the transceiver/signal processor and the 9 track tape drive. Located on top of the desk is the ARTIST ionogram scaler and its associated printer and modem.

50, 100 or 200 kHz. The lower and upper frequency limits are 0.5 and 30.0 MHz. Although logarithmic frequency scans are available, they are rarely used and most systems operate with 100 kHz increments for vertical incidence ionograms. The time dwell on each frequency determines the Doppler resolution of the discrete Fourier Transform, and must be appropriately selected for different ionospheric conditions. Typical values for the dwell lie between 0.5 and 2.0 seconds, resulting in Doppler resolutions of 2 to 0.5 Hz. The Digisonde ionograms have either 128 or 256 height ranges, and for each range a 16 point Doppler spectrum is calculated (8 points only for 256 height ranges). It is difficult to visually present an ionogram that is composed of 128 spectra for each sounding frequency. The situation becomes more complicated when incidence angle and signal polarization are considered. To get an indication where the echoes in the ionogram are coming from, the antenna beam of the receiving array is steered from vertical (where it receives both in O and X polarization) to several off zenith directions. Up to 12 azimuthal directions can be scanned for each of two zenith angles, 10° or 21°. The different directions and the O/X switching are time sequenced at the pulse repetition rate, and only one receiver is used.

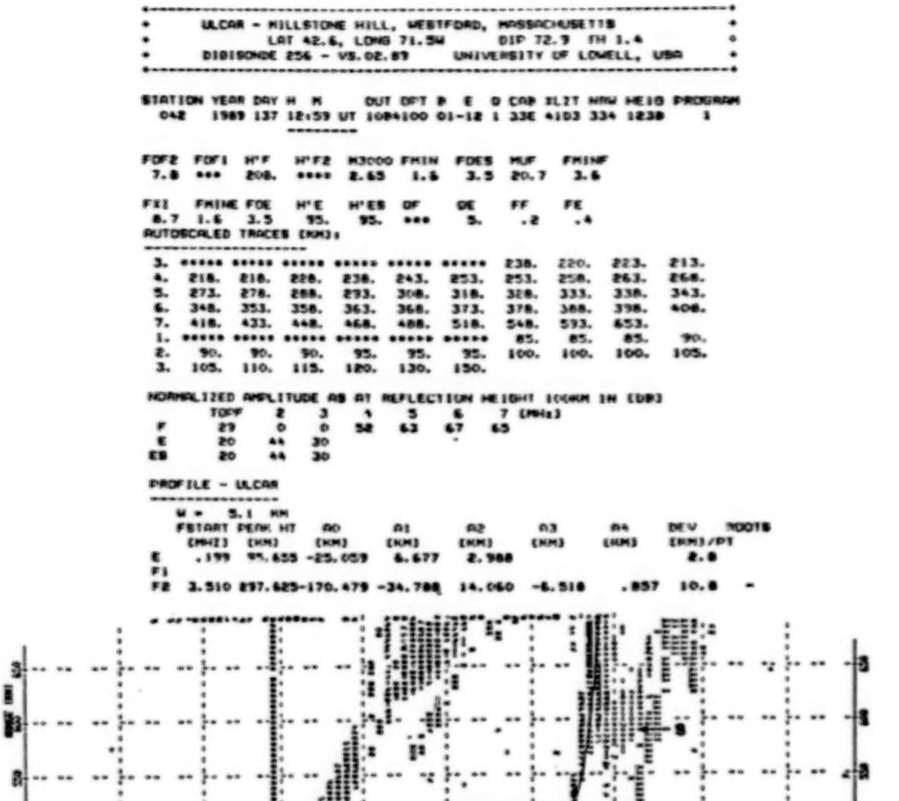
The decision was made, somewhat arbitrarily, to limit the number of observations or "channels" to 16, e.g. 16 Doppler lines for the ordinary polarized signals, or eight Doppler lines each for O and X polarization, or just two Dopplers each for six incidence angles and the vertical O and X channels. To preserve the full information content all 16 channels must be recorded on magnetic tape, and one data block with 4096 bytes per sounding frequency is written onto the 1600 bpi tape. Generally this is not done, however, since a 2400' tape can only store about 50 of these full information ionograms. Instead, the 16 channels for each frequency/range bin are analyzed, the channel with the maximum amplitude is selected, and the amplitude and channel number, or "status", is recorded. This technique makes it also possible to present the ionogram in the conventional frequency versus range frame, as illustrated in Figure 3.

The sequence of soundings is freely selectable and the system can be preprogrammed to change to a different sequence at any desired day and time. For example, many stations change from the routine 15 min ionogram/drift sequence to a 5 min sequence for the Regular World Days. Different ionogram programs can be preprogrammed and the system will sequence through them as well as through fixed frequency or drift observations. The printout uses an optically weighted font [Bibl and Reinisch, 1978] to present the amplitudes of the overhead ordinary echo traces; the letter X represents the extraordinary traces, and little arrows point to the direction from where oblique echoes are received. On the Millstone Hill (42°N geogr.) ionogram in Figure 3, F region echoes from the south were received at around 8 MHz, and Es echoes from north-east at 3 MHz. The lettering within the ionogram was added for this publication.

The fixed frequency mode is identical to the ionogram mode, except the sounding repeats on the same frequency a selectable number of times. It is also possible to step through four different frequencies, and repeat these transmissions n times. Recording format and the printouts are the same as in the ionogram mode. The drift mode, on the other hand, is very different in its operation and is explained in some detail in Section 5.

Signal Processing

The received RF signals are amplified in tuned circuits, enabling successful signal reception in regions with high HF communication activities, and converted to an intermediate frequency (IF) of 225 kHz. A 12 bit digitizer takes quadrature samples of the linear IF signal. In the ionogram mode 128 or 256 pairs of quadrature samples are taken which are spaced by 16.66, 33.33, or 66.66 μsec corresponding to 2.5, 5.0, or 10.0 km height resolution, and resulting in range coverages varying from 320 to 1280 km. By using the phase difference between simultaneously transmitted signals, offset in frequency by 5, 15 or 30 kHz, a group path accuracy of better than 1 km can be obtained. For high pulse repetition rates (200, 100, and 50 Hz are available) higher order echoes can fall into the digitizing windows. However, the bi-phase pulse sequence modulation, impressed onto the transmitted signal, results in significant suppression of these high-order echoes during the phase coherent spectral integration process. The symmetric bi-phase sequence code has been optimized for all different ionogram and fixed-frequency programs so as to



```
*-----------------------------------------------------------------*
*      ULCAR - MILLSTONE HILL, WESTFORD, MASSACHUSETTS            *
*            LAT 42.6, LONG 71.5W       DIP 72.7  FH 1.4          *
*      DIGISONDE 256 - V5.02.89      UNIVERSITY OF LOWELL, USA    *
*-----------------------------------------------------------------*

STATION YEAR DAY H  M      OUT OPT P  E  D CAP XLIT HAW HEIG PROGRAM
  042   1989 137 12:59 UT 1084100 01-12 1 33E 4103 334 1238      1
                 --------

FOF2  FOF1  H'F   H'F2  M3000 FMIN  FOES  MUF   FMINF
 7.8  ***   208.  ****  2.65   1.6   3.5  20.7   3.6

FXI   FMINE FOE   H'E   H'ES  OF    OE    FF    FE
 8.7  1.6   3.5   95.   95.   ***   5.    .2    .4
AUTOSCALED TRACES (KM):
---------------------
 3. ***** ***** ***** ***** ***** *****  238.  220.  223.  213.
 4.  218.  218.  228.  238.  243.  253.  253.  258.  263.  268.
 5.  273.  278.  288.  293.  308.  318.  328.  333.  338.  343.
 6.  348.  353.  358.  363.  368.  373.  378.  388.  398.  408.
 7.  418.  433.  448.  468.  488.  518.  548.  593.  653.
 1. ***** ***** ***** ***** ***** *****   85.   85.   85.   90.
 2.   90.   90.   90.   95.   95.   95.  100.  100.  100.  105.
 3.  105.  110.  115.  120.  130.  150.

NORMALIZED AMPLITUDE AS AT REFLECTION HEIGHT (KM) IN (DB)
       TOFF   2   3   4   5   6   7 (MHz)
 F      29    0   0  58  63  67  65
 E      20   44  30
 ES     20   44  30

PROFILE - ULCAR
---------------
   W =   5.1  KM
   FSTART PEAK HT   A0       A1       A2       A3       A4      DEV    ROOTS
    (MHZ)  (KM)     (KM)     (KM)     (KM)     (KM)     (KM)    (KM)/PT
 E   .199  95.655 -25.059   6.677    2.988                        2.8
 F1
 F2 3.510 237.625-170.479 -34.788   14.060   -6.518    .857     10.8    -
```

Figure 3 -- The Digisonde 256 on-line ionogram printout includes (from top to bottom) information on the station of observation, system settings, scaled ionospheric parameters, autoscaled trace points, reflection amplitude information and electron density profile parameters. The ionogram is displayed in the lower half of the figure and indicates the ordinary and extraordinary traces, oblique signals with their directions (S, NE) and the calculated electron density profile. Lettering and lines were added for publication. This data is from Millstone Hill, MA at 07:59 LT on 17 May 1989.

suppress these multiples as well as coherent internal noise and external interferers. To avoid the strongest interferers, the Digisonde employs a frequency search technique. Prior to transmission of a frequency f_o, the Digisonde listens at frequencies f_o, f_o + 10 kHz, and f_o + 20 kHz and uses the one with the lowest interference level as sounding frequency.

A 20-bit discrete Fourier transform calculates the complex spectrum (amplitude and phase) for all height ranges in real time. In the ionogram mode 16 spectral lines are calculated, less when polarization and angle of arrival are measured. In the drift mode up to 256 spectral lines for a limited number of range gates are determined. The spectral data are outputted onto tape and paper; the available amplitude and phase resolution is 0.25 dB and 1.4° respectively.

To handle the large range in signal amplitudes which vary with frequency and range and are controlled by focussing effects, diurnal variation in absorption and geomagnetic activities, the dynamic range of the receiver system has been enlarged to 100 dB by using time-dependent (day/night) and signal-dependent digital gain control. Coherent integration of echoes from three pulses transmitted prior to the actual sample pulses sets the optimum gain level for the integration period thus keeping the gain constant during the coherent spectral integration.

Oblique Ionogram Sounding

Oblique sounding studies between Digisondes separated by distances of 350 to 5000 km have been conducted [Reinisch et al., 1984]. These soundings can be made either in the ionogram or the fixed-frequency mode. Figures 4 and 5 show examples from recent experiments between Goose Bay and Millstone Hill, a distance of 1400 km. In Figure 4 the sequence of afternoon ionograms on 24 August 1989 starting at 1420 LT (1920 UT) indicates a build-up of the F2 propagation mode and the decline of the F1 ledge. The late evening sequence in Figure 5 shows the decay of the F layer. An automatic scaling algorithm for oblique ionograms over distances of 3000 km or less is currently being developed and will result in a quasi-parabolic description of the mid-point vertical electron density profile [Kuklinski et al., 1988].

The internal timing of the Digisonde is controlled by timing pulses that are derived from the system oscillator, which for bistatic sounding is a rubidium standard with a stability of 1 part in 10^{11}. The clocks in the two Digisondes must be synchronized with an external one second pulse to make oblique ionograms. We are currently testing an automatic time synchronization technique that locks the Digisonde clock within 10 µs to an external time reference such as the GOES or GPS satellite.

Automatic Scaling of Ionograms

The scaling of ionograms has always been a time consuming process that required well trained personnel. A fairly detailed understanding of ionospheric radiowave propagation is required to properly interpret the ionograms. Even the experts have often difficulty in analyzing ionograms in the presence of spread F and oblique signals. This task has become much easier with Digisonde ionograms that identify each echo pixel as either vertical or off vertical, as either O or X polarization. Indeed, now that these signal characteristics are measured it has become possible to automate the routine scaling of ionograms, using an expert system. In the Digisonde 256 the ARTIST software [Reinisch and Huang, 1983; Tang et al., 1988], running on an IBM AT or clone, determines the leading edges of the ordinary echo traces of the E, Es, F1 and F2 layers, scales the standard ionogram characteristics, shown at the top of Figure 3, and calculates the vertical electron density profile [Huang and Reinisch, 1982] as described in Section 4 of this paper. Depending on foF2, the scaling and true height calculation is generally completed within 20 to 40 sec after the ionogram scan has ended. Currently, ARTIST uses the Microsoft MS DOS 3.2 operating system and is written in FORTRAN 4.01. The US Air Weather Service network, called DISS (Digital Ionospheric Sounding System) still maintains the older software of 1987 (MS.DOS 3.2, Microsoft FORTRAN 3.31). For every routine ionogram, ARTIST scales the following characteristics: foF2, foF1, h'F, h'F2, M3000, fmin, foEs, MUF(3000), fminF, fxI, fminE, foE,

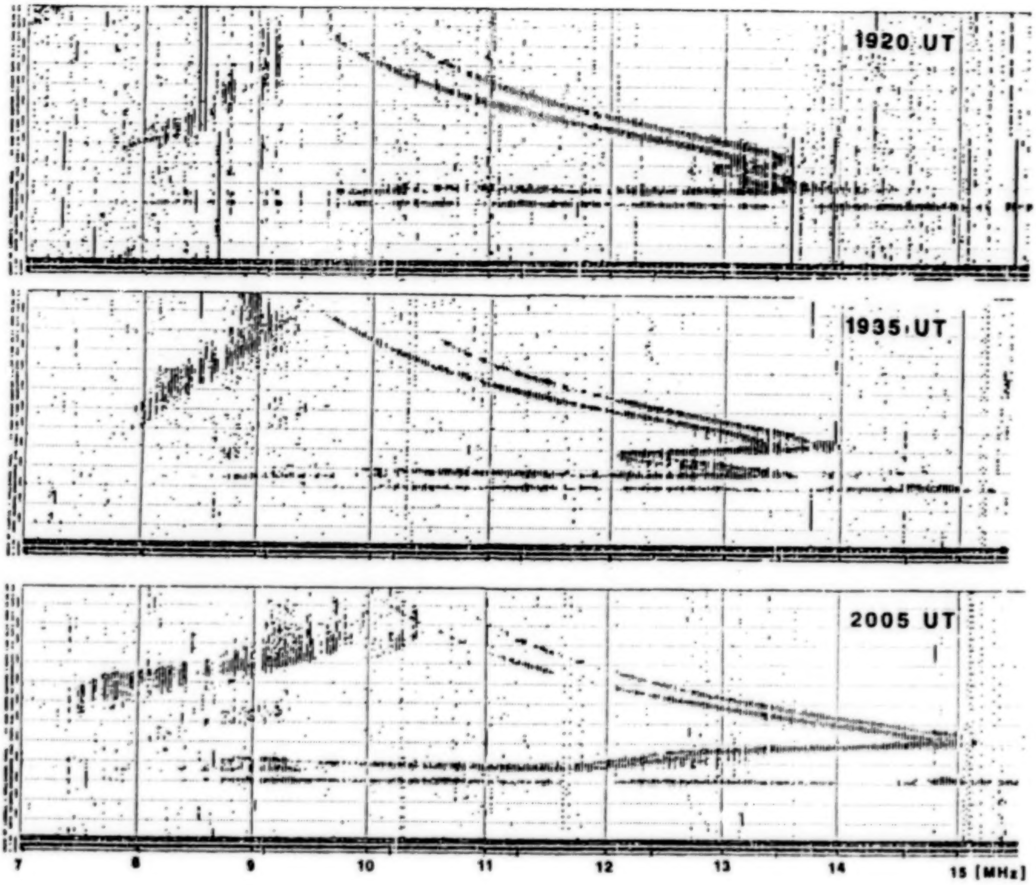


Figure 4 -- Oblique bistatic ionograms from Goose Bay, Labrador to Millstone Hill, MA. This sequence of 3 ionograms was made during the afternoon (19:20-20:05 UT) on 24 August 1989. Single hop E, F1 and F2 as well as two hop F2 reflections are present. The time sequence shows the decline of the F1 layer and an increase of the F2 layer.

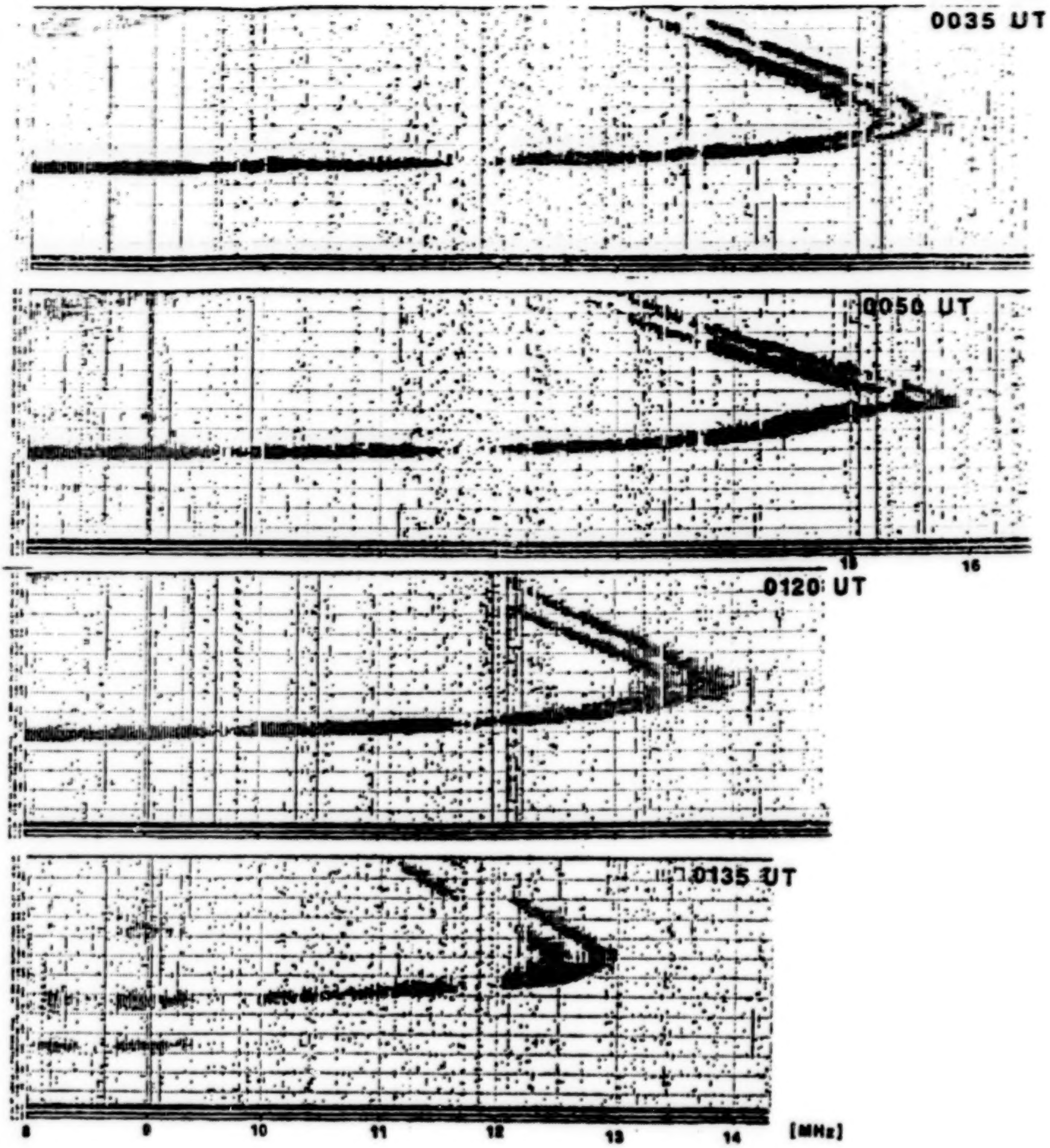


Figure 5 -- Oblique bistatic ionograms from Goose Bay, Labrador to Millstone Hill, MA. This sequence of 4 ionograms was made during the early evening (00:35-01:35 UT) on 24 August 1989. This figure depicts the change in the nighttime F2 region propagation as the evening progresses.

h'E, h'Es, and the average range and frequency spread for E and F traces. The virtual heights of the E and F region echo traces are recorded in km for every 100 kHz frequency increment, together with the echo amplitude and Doppler shift.

The automatic scaling of ionograms makes it feasible to operate at higher data rates, like four ionograms per hour or 12 per hour. The advantage of higher data rates is twofold. For once, fast changes in the ionosphere can be monitored, and secondly there is some data redundancy so that occasional scaling errors have much less weight than in the case of hourly sounding. For the same reason the so-called descriptive and qualifying letters [Piggott and Rawer, 1972] are losing in importance. The high rate data sequence, supplemented by amplitude, Doppler and spread information, generally gives a better description of the ionosphere than could be obtained by hourly readings with descriptive/qualifying letters. The definition of these letters was not made for automatic scaling, and ARTIST does not use any letters. URSI Commission G Working Group G4 during a July 1989 workshop at Lowell, MA, USA, has recommended the data format for automatically scaled ionogram data shown in Table 1.

While autoscaling is a very efficient way of making the most basic ionogram information available for geophysical research it is prone to scaling errors. This issue is addressed in Section 4. Work to further improve the autoscaling is in progress. Recent modifications of the ARTIST algorithms [Tang et al., 1988] have improved the scaling of difficult ionograms. The old ARTIST often failed in the presence of strong multiple Es traces as in the case of the mid-latitude ionogram in Figure 6, where a wrong F trace was identified to start at 1.9 MHz. The new program identifies the F trace correctly with fminF = 4.4 MHz. In the polar cap, the autoscaling had occasionally latched on to improperly identified oblique echoes (Figure 7), the new program avoids this mistake.

Remote Sounder Control and Data Output

One convenient operational feature for scientific applications is the capability to remotely change the operating parameters and the sounding schedule over dial-up phone lines, and to use the same lines for transfer of ionogram data. The Digisonde's ARTIST computer provides two full-featured serial RS232 ports for remote communication at 300, 1200, 1200/75, and 2400 baud via low cost modems and voice grade telephone lines. Most Digisondes are connected to a 2400 baud dial-up modem. Frequently the other full-featured port is connected over dedicated lines to a local terminal at the same facility in another building. In the case of Argentia, Newfoundland, the dedicated line connects the unmanned station to a terminal at the Goose Bay Ionospheric Observatory, some 900 km away. A third serial port is designed to accommodate the special polling procedure of the U.S. Air Weather Service, and to put the ARTIST scaled parameters, trace coordinates, and true height coefficients into a fixed format message. Other users have connected this port or one of the two full-featured ports to another computer for either an automated control function or a remote data gathering application; no additional hardware is required.

The ULCAR Remote Terminal is an IBM PC clone using the same display adapter, monitor, communication hardware, printer and the same software used by the ARTIST for display, printing and communications. The Remote Terminal is connected to one of the two full featured ports over dial-up modems or dedicated lines. Operation of the Digisonde from a Remote Terminal is virtually identical to the onsite operation control. The data output capability is presently limited to the processed ionogram data in the format shown in Figure 3. To make the "raw" 16-channel ionogram and the drift mode data available at the Remote Terminal, requires development of additional software for the ARTIST and the Remote Terminal. Currently available polling features allow the remote user to obtain the last complete ionogram in the format shown in Figure 3, and also the last 24 hourly values of foF2, foF1, and MUF3000. The data base available for polling could easily be expanded.

At the present time eleven Digisondes have dial-up modems attached. Access is generally restricted to those with a direct involvement in the station operation but in principle, access could be made available to the broader scientific community. Addition of operational security software is expected

Code	Format	Description
	80I3	DATA FILE INDEX
1	10F7.3	GEOPHYSICAL CONSTANTS
2	120Z1	IONOGRAM SOUNDING SETTINGS (PREFACE)
3	50F8.3	SCALED IONOSPHERIC PARAMETERS
4	20I2	ARTIST ANALYSIS FLAGS
5	16F7.3	DOPPLER TRANSLATION TABLE
		O-TRACE POINTS - F2 LAYER
6	400I3	VIRTUAL HEIGHTS
7	400I3	TRUE HEIGHTS
8	400I2	AMPLITUDES
9	400I1	DOPPLER NUMBER
10	400F6.3	FREQUENCY TABLE
		O-TRACE POINTS - F1 LAYER
11	150I3	VIRTUAL HEIGHTS
12	400I3	TRUE HEIGHTS
13	150I2	AMPLITUDES
14	150I1	DOPPLER NUMBER
15	150F6.3	FREQUENCY TABLE
		O-TRACE POINTS - E LAYER
16	150I3	VIRTUAL HEIGHTS
17	400I3	TRUE HEIGHTS
18	150I2	AMPLITUDES
19	150I1	DOPPLER NUMBER
20	150F6.3	FREQUENCY TABLE
		X-TRACE POINTS - F2 LAYER
21	400I3	VIRTUAL HEIGHTS
22	400I2	AMPLITUDES
23	400I1	DOPPLER NUMBER
24	400F6.3	FREQUENCY TABLE
		X-TRACE POINTS - F1 LAYER
25	150I3	VIRTUAL HEIGHTS
26	150I2	AMPLITUDES
27	150I1	DOPPLER NUMBER
28	150F6.3	FREQUENCY TABLE
		X-TRACE POINTS - E LAYER
29	150I3	VIRTUAL HEIGHTS
30	150I2	AMPLITUDES
31	150I1	DOPPLER NUMBER
32	150F6.3	FREQUENCY TABLE
33	20I2	MEDIAN AMPLITUDE OF F ECHO
34	20I2	MEDIAN AMPLITUDE OF E ECHO
35	20I2	MEDIAN AMPLITUDE OF ES ECHO
36	20E9.4E1	TRUE HEIGHT F2 LAYER COEFFICIENTS
37	20E9.4E1	TRUE HEIGHT F1 LAYER COEFFICIENTS
38	20E9.4E1	TRUE HEIGHT E LAYER COEFFICIENTS
39	20E9.4E1	TRUE HEIGHT MONOTONIC SOLUTION
40	20E9.4E1	VALLEY DESCRIPTION

FORMAT FOR AUTOMATICALLY SCALED IONOGRAM DATA

RECOMMENDED BY URSI WORKING GROUP G4 ON
IONOSPHERIC INFORMATICS

TABLE 1

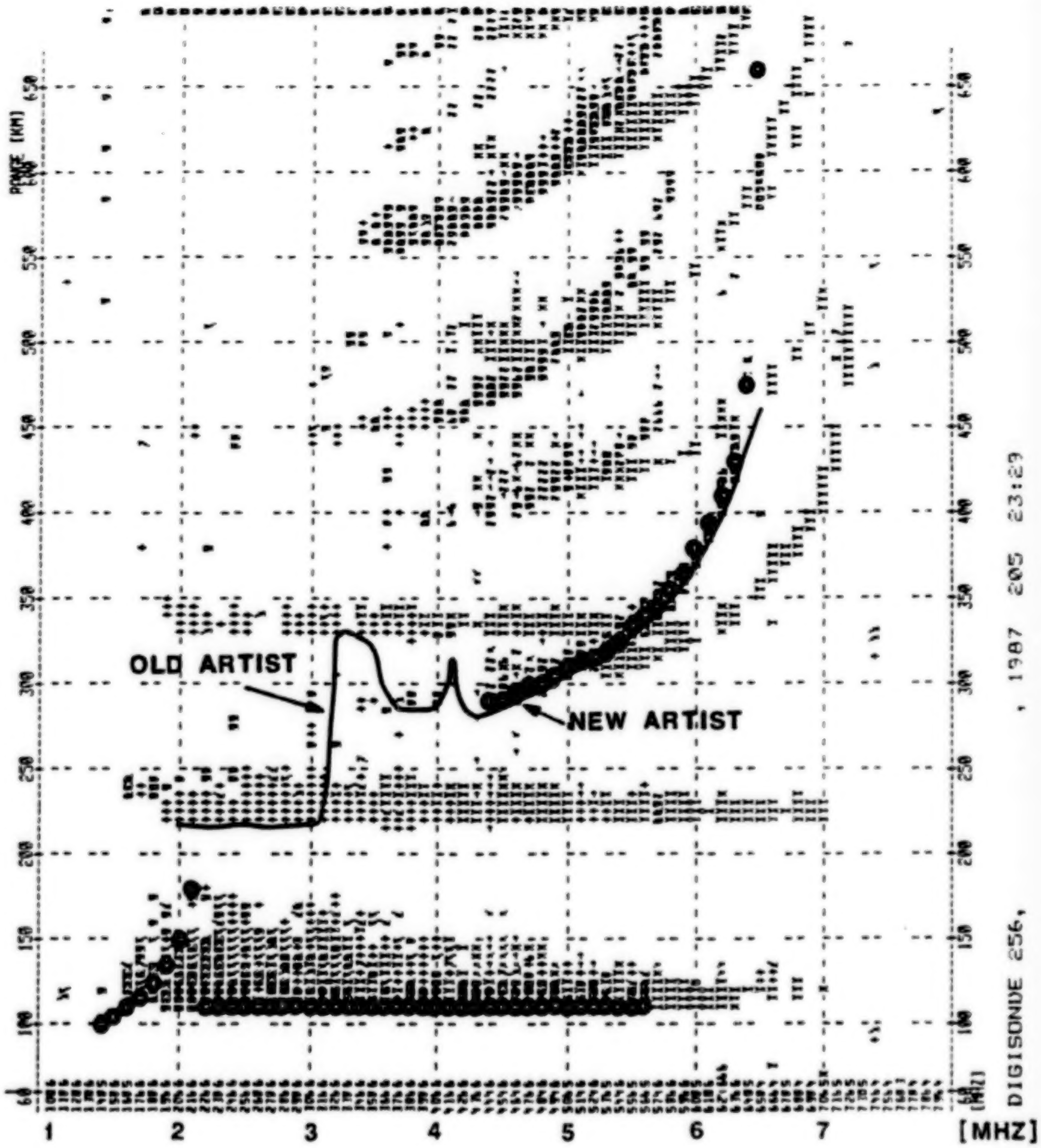


Figure 6 -- ARTIST results indicating the improvement in automatic scaling at mid-latitude stations in the presence of multiple sporadic E reflections. This data is from Millstone Hill at 19:29 LT on 24 July 1987

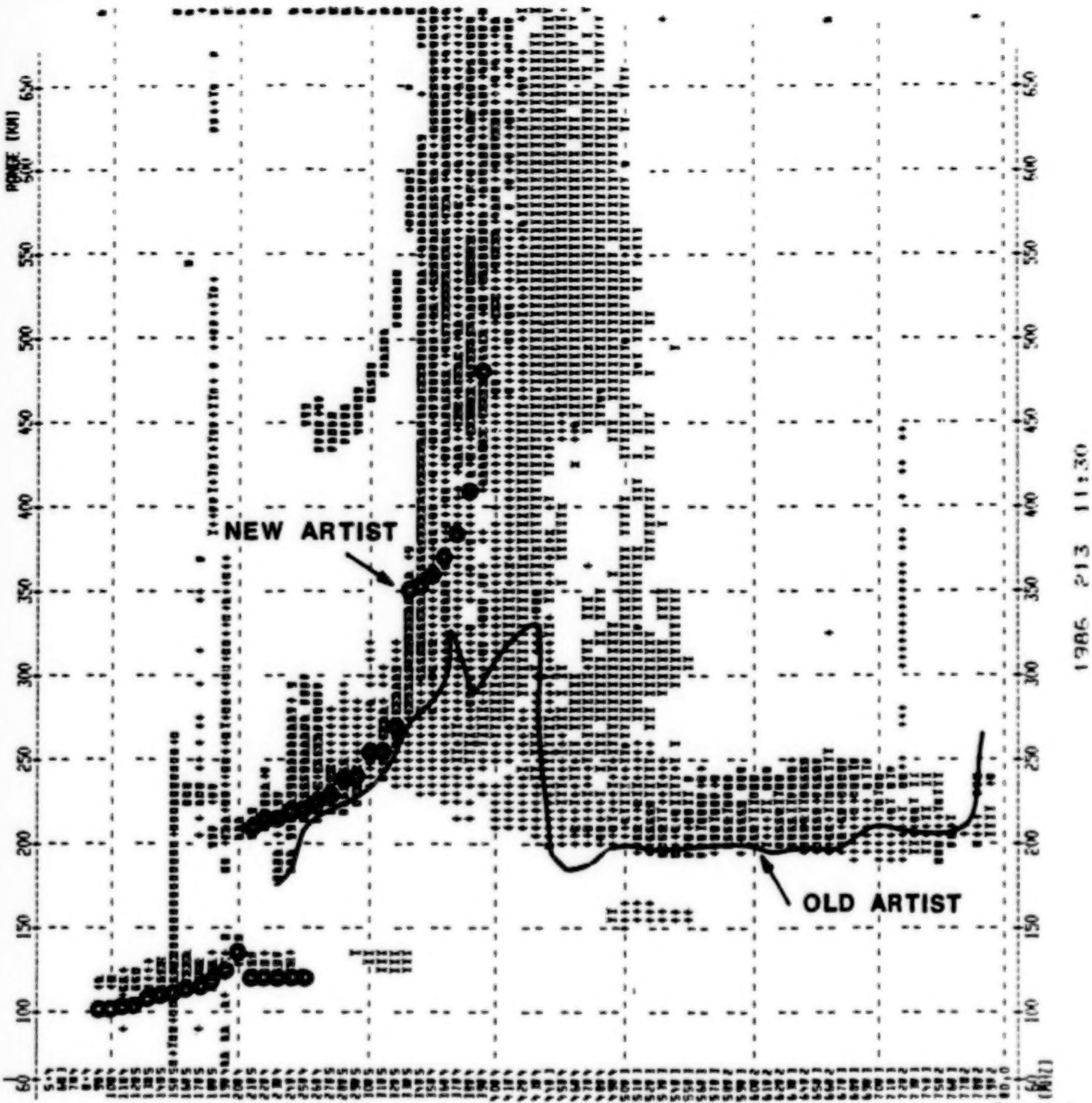


Figure 7 -- ARTIST results indicating the improvement in automatic scaling at high latitude stations during spread-F conditions. This data is from Qaanaaq, Greenland at 14:30 LT on 01 August 1986.

for the near future to protect routine operation of the Digisondes from inadvertent interruption; this will enable broader access to the Digisonde network.

3. DESKTOP PROCESSING AND DISPLAY OF IONOGRAM DATA

Ionograms from analog ionosondes were normally recorded on black and white film and no further processing need be done (nor could it be done) to view these data. Furthermore, the information content of these analog ionograms is low, consisting of a 1-bit echo/no echo indication for each frequency/range bin. The Digisonde ionograms, on the other hand, are available on magnetic tape in computer-readable form, containing the multiparameter information described in Section 2. Software for a desktop computer (an amended IBM AT, configured like the on-line ARTIST equipment, with added color display and color ink jet printer) has been developed which makes it possible to display, edit, and print the multiparameter ionograms, scaled characteristics, and the electron density profiles as function of time.

This ionospheric data management and analysis software system, called ADEP (for ARTIST Data Editing and Printing),reads the ionogram data and ARTIST-scaled ionospheric characteristics from magnetic tape, and displays the digital ionogram along with the scaled traces and characteristics on a color monitor for the purpose of data analysis or rescaling. Any scaling errors in the autoscaled traces or characteristics can be corrected before the data are written to an archive data file (Table 1) for further study or global data exchange. This "editing" of the ARTIST data can be performed on an ionogram-by-ionogram basis or in a search mode where the magnetic tape is stopped only when specified inconsistencies in the data sequence are detected.

To show the diurnal variations in the ionosphere ADEP plots all quarter-hour ionograms for an entire day on one sheet as illustrated in Figure 8 for Millstone Hill on 20 February 1988. In this example the extraordinary and oblique echoes are suppressed giving a clearer view of the overhead plasma frequency and height variations. These ionogram surveys are helpful in identifying absorption events and the passage of irregularities. The onsets of the high latitude F-region trough and the auroral oval can easily be detected. F-region patches [Buchau et al., 1983], driven across the polar cap by the ExB drift, show as sudden increases in foF2 in the sequence of ionograms.

For ionosphere/thermosphere studies the coupling between different height regions is better seen on the time evolution of the electron density distributions than on the sequence of ionograms. ADEP uses the ARTIST profiles to generate contour plots for 24 hour periods like the example in Figure 9. These contour plots contain the most important information obtained from a day's worth of ionograms. The isodensity contours are represented by contour lines with 0.5 MHz increments in plasma frequency. The data are plotted versus universal time (UT), and local midnight (M), noon (N) and sunrise/sunset in the F and E region (F and E) are marked at the bottom and on top. The peak height of the F layer, hmF, is plotted as a dotted curve. The row of double dots in the lower part of the figure, labeled DATA AVAILIBLITY, indicates the times when ionograms were recorded (upper dot) and whether an electron density profile was calculated (lower dot). In the upper part of the figure the critical frequencies foF2, foF1, foE and foEs are plotted. The vertical lines on the curves specify the extend of the measured frequency spread.

For modeling and predictions, monthly data displays are more suitable than the daily presentations, since they reveal the day-to-day variations much quicker.. An example for a monthly display is given in Figure 10. Here the deviations of foF2 from the quiet-day median is plotted for Argentia, Newfoundland, December 1986. For reference, the Kp indices, obtained from World Data Center A, are plotted above the foF2 data.

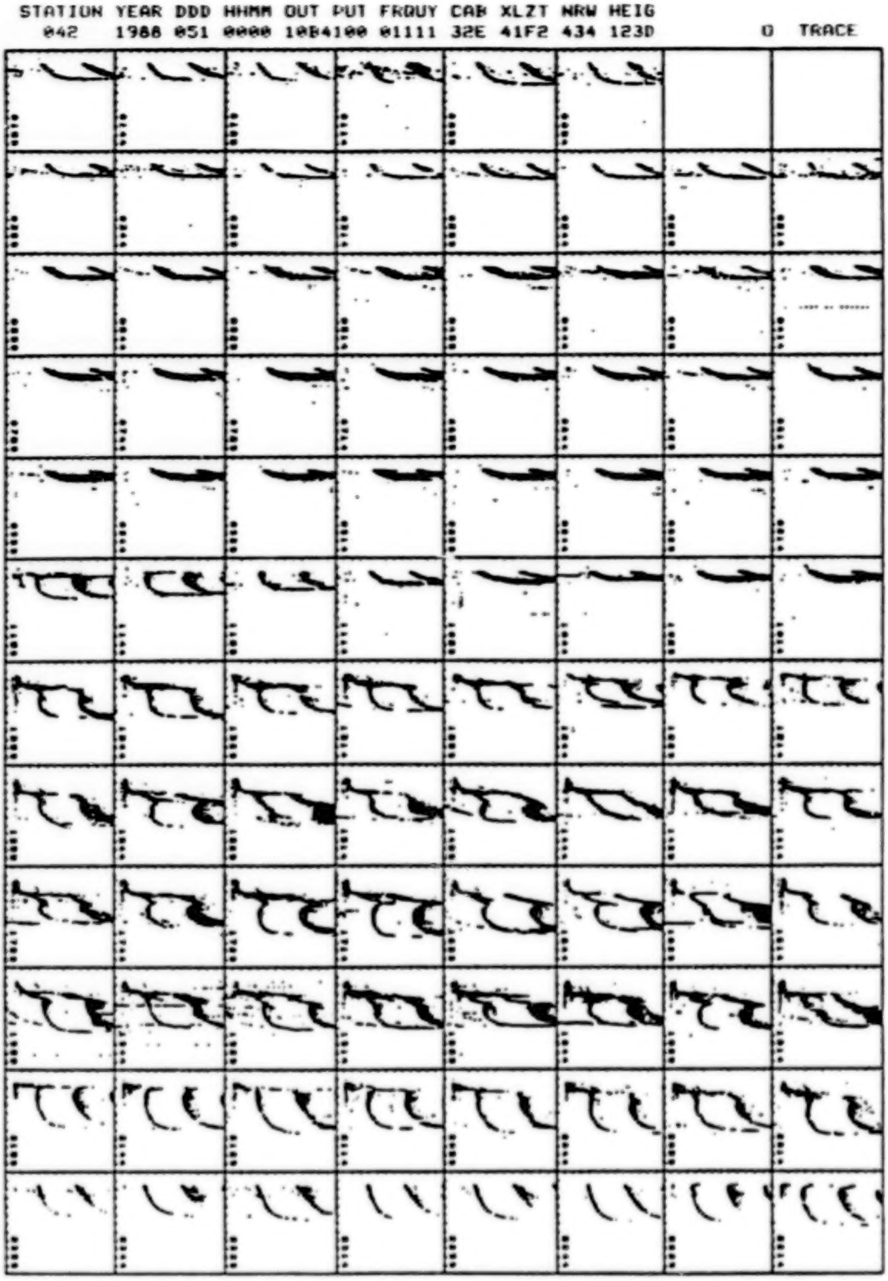


Figure 8 -- Ionogram survey for 20 February 1988 from Millstone Hill, MA. The ordinary component is displayed while the extraordinary and oblique signals are suppressed. Ionograms are shown at 15 minute intervals over a 24 hour period from 00:00 UT to 23:59 UT (19:00 LT to 18:59 LT).

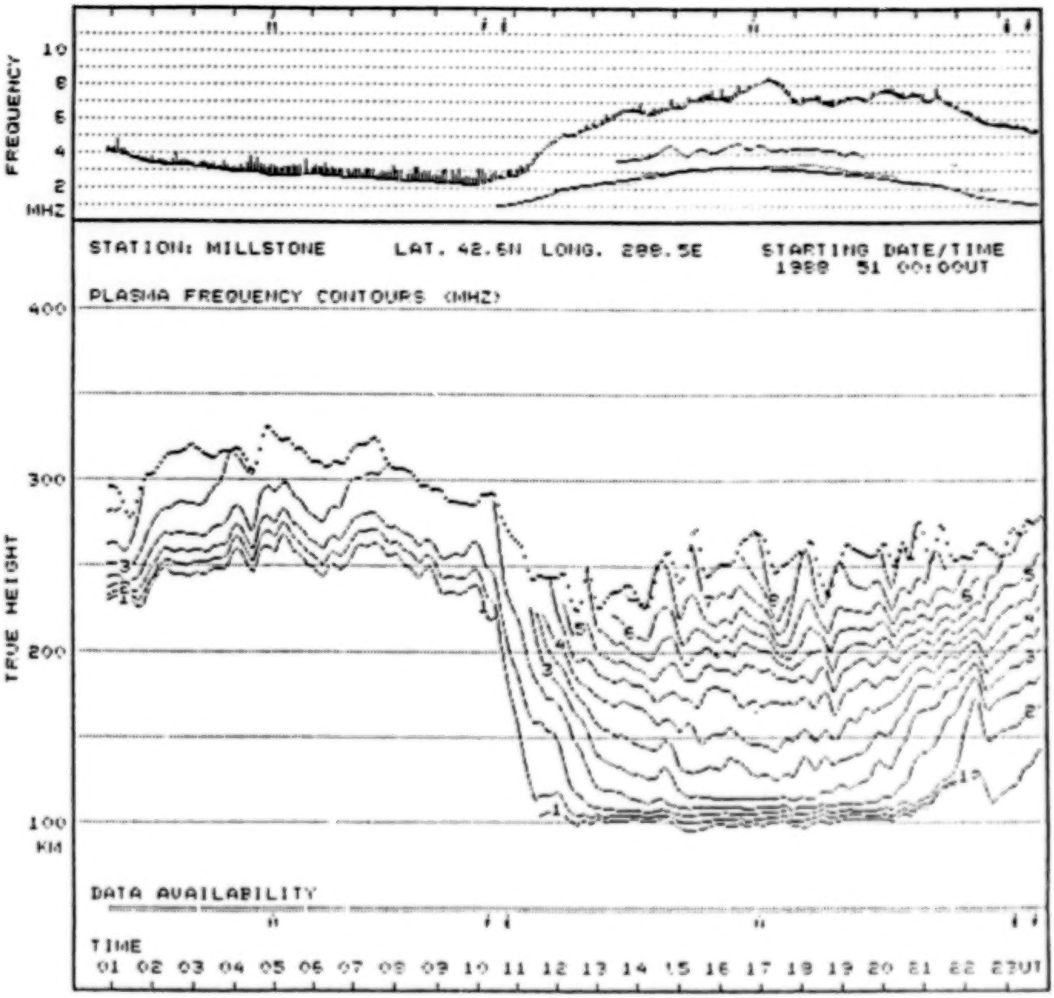


Figure 9 -- Plasma frequency contours vs time for 20 February 1988 from Millstone Hill, MA. The top panel indicates foF2, foF1, foE and foEs over the 24 hour period from 00:00 UT to 23:59 UT. The bottom panel shows plasma frequency contours in true height and time. Contour lines are spaces at 0.5 MHz intervals. The data availability line indicates the presence of data as well as local noon, midnight, and sunrise/sunset in the E and F regions.

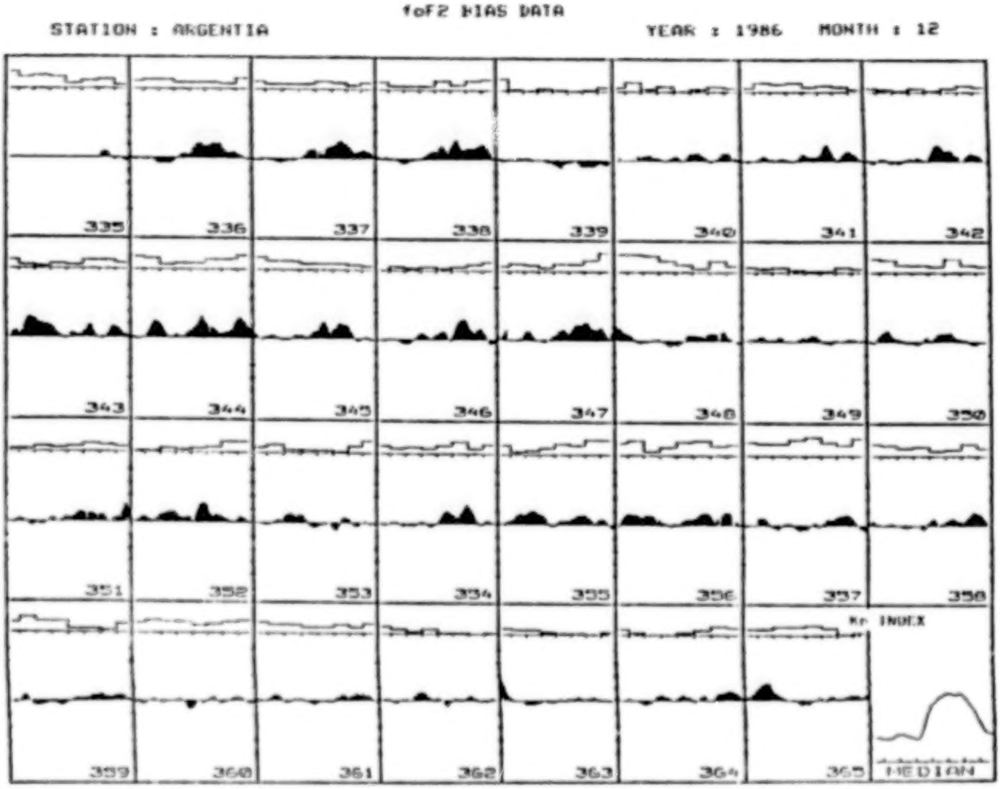


Figure 10 -- Daily display of foF2 deviations from the monthly median for December, 1986. This data is from Argentia, Nfld. Plotted above each daily foF2 deviation is the Kp index for that day. The last panel in the lower right shows the monthly median values (for quiet days) from which the deviation is calculated.

4. REAL TIME ELECTRON DENSITY PROFILES

ARTIST directly converts the autoscaled traces into a vertical electron density profile; the time required is approximately 20 sec. In the inversion process, the electron density profile in each ionospheric region (E, F1 and F2) is given as [Reinisch et al., 1988; Huang and Reinisch, 1982]

$$Z(f) - Z(f_s) = A_{I+1} + g^{1/2} \sum_{i=0}^{I} A_i T_i^*(g) \tag{1}$$

$$g = \frac{\ln f/fc}{\ln fs/fc} \qquad 0 \le g \le 1$$

f = plasma frequency,

f_c = critical frequency,

f_s = starting frequency

T^*_i = shifted Chebyshev polynomial [Snyder, 1966]

A_{I+1} = layer half thickness, $Z(f_c) - Z(f_s)$,

$= -\sum_{i=0}^{I} A_i$; I = 4 for F region, I = 2 for E region.

The coefficients A_i are determined in a least squares fit by minimizing the difference between the $h'_{cal}(f)$ function calculated from equation (1) and the measured h'(f) curve. Huang [1989] has shown that equation (1) is an approximation to a Chapman profile with varying scale height. For the routine calculation of real time profiles a standard POLAN valley is used as proposed by Titheridge [1985], where the valley width W and depth D are a function of the peak height of the E-layer [Reinisch et al., 1989].

Use of a model valley introduces errors in the height of the daytime F layer, especially for plasma frequencies just above foE, the critical frequency of the E layer, when the actual valley shape differs significantly from the model. However, it is often more important to see the diurnal and seasonal variations of the electron density distribution and the effects of solar events and magnetic storms on it than to specify the height with a 10-20 km accuracy. To assess the coupling of the magnetosphere to the F2 ionization and of the mesosphere/thermosphere to the E region, electron contour plots like the one in Figure 9 become invaluable when they are generated for the entire Digisonde network.

Ongoing research tries to resolve the valley ambiguity caused by the fact that no HF waves are reflected from the valley between the E and F regions. Xu et al. [1989] use the change in echo amplitudes for frequencies below and above foE to estimate the valley width W and the depth D. The difference in non-deviate absorption between signals reflected in the F ($f \ge foE$) and the E layer ($f \le foE$) is equal to

$$\Delta L = \frac{e^2}{4\pi^2 mc} \int_{hE}^{hF} \frac{N\nu}{(foE + f_L)^2}\, dh \,[\text{nepers}]$$

where e, m, N are the electron charge, mass and density, ν is the effective collision frequency assumed to vary exponentially with height over the valley region, and f_L is the longitudinal component of the gyrofrequency. By measuring ΔL, an estimate can be made of the electron distribution in the valley region $hE < h < hF$. We are currently investigating an iterative approach to determine both hF, the height of the first F reflection for $f \geq foE$, and the valley ionization N_V.

Initial work has started with Digisonde data from Millstone Hill and the calculated profiles are compared with the profiles obtained with the Millstone Hill incoherent scatter radar [Buonsanto, 1989]. The initial differences between the ISR radar and the Digisonde profiles calculated with the standard POLAN valley are illustrated in Figure 11 which shows the lower portion of the noontime electron density profile over Millstone Hill for 10 December 1988. The ISR profile was calibrated with the the foE and hmE values from the Digisonde, so the profiles agree well up to hmE. But the F layer heights of the Digisonde profile are too low by about 15 km since the standard (model) valley is too shallow. Since ARTIST outputs the complete h'(f) traces it is always possible to recalculate the profiles using different inversion algorithms.

5. DIGISONDE DRIFT OBSERVATIONS

Spaced-antenna HF observations of ionospheric drifts is an effective method of studying the structure and the dynamics of the ionosphere. In the drift mode the Digisonde operates at 1, 2 or 4 frequencies and records the spectra of the signals received at four or seven receiving antennas. The principal idea [Pfister, 1974] of this drift technique is similar to that of a multibeam radar. The widebeam transmit antenna illuminates an area of several hundred kilometer diameter in the ionosphere. Echoes will be returned to the sounder from all points on the isodensity surface $f_N = f$, where the surface vector points toward the sounder (neglecting refraction effects). In the presence of any undulations or irregularities more than one reflection point exists; and since there are generally many more than three reflection points, each with a distinct Doppler shift, one can construct the three-dimensional drift velocity vector for the ionospheric motion when the following conditions are satisfied: one, the drift velocity is uniform over the illuminated area, and two, refraction effects can be neglected.

It is important to realize that closely spaced reflection points in the ionosphere cannot be separated with interferometric techniques considering the receiving array dimensions of typically 100 m. For a 10 MHz signal, the 6 dB beam width is about 15°, and correspondingly larger for lower frequencies. To circumvent this difficulty, the Digisonde Fourier analyzes each individual antenna signal [Bibl and Reinisch, 1978], thus separating the different arriving waves in the spectral domain. The discrete Fourier transform determines the complex spectrum (A_{ij}, ϕ_{ij}), with $j = 1, \ldots, 128$ (128 spectral lines) and $i = 1, \ldots, 4$ or 7 (for four or seven receiving antennas). The spectral index j gives the Doppler frequency d. The echo arriving from source s has a Doppler shift.

$$d_s = \frac{1}{\pi} \mathbf{v} \bullet \mathbf{k}_s , \ s = 1, 2, \ldots \text{ (source number)} \tag{2}$$

where $\mathbf{v}$ is the drift velocity and $\mathbf{k_s}$ the wave vector pointing from reflection point s to the receiving antennas. From the spectral phases Ø_{is}, the wave vector $\mathbf{k_s}$ is determined, i.e. the source (reflection point) of the wave with Doppler d_s. The intersection of the vectors $-\mathbf{k_s}$ with the horizontal plane at the altitude of the radar range (height gate) of the observation generates the so-called skymaps [Bibl et al., 1975].

Figure 12 shows a sequence of four consecutive skymaps, the overhead position is in the center of the map and zenith angles up to 40° are indicated by tick marks at 5° intervals. These measurements were made at Millstone Hill on 11 March 1989 with the sounding frequency f = 3.125 MHz. Each skymap is the result of a 20 second integration, corresponding to a Doppler

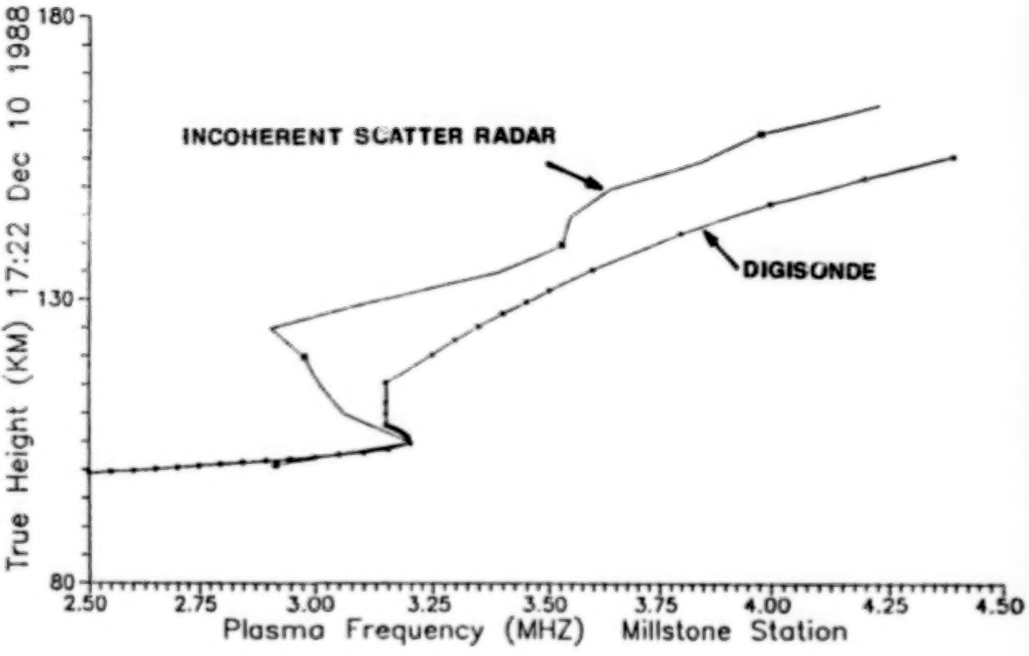


Figure 11 -- Comparison of electron density profiles in the E and lower F regions as measured by Digisonde and incoherent scatter radar. Agreement is quite close up to foE but the E region valley uncertainties cause discrepancies in the F region heights. This data is from Millstone Hill at 12:22 LT on 10 December 1988.

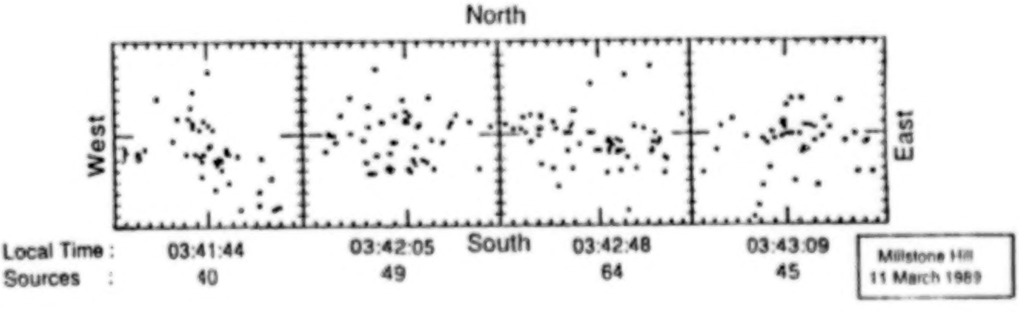


Figure 12 -- Skymap sequence from Millstone Hill depicting sources located during four successive 20 second long measurements. Local time and total number of sources in each skymap are shown. Each point represents a wave component of significant amplitude whose location indicates the arrival angle (azimuth and zenith) of a reflected signal. The center of the skymap is overhead while the middle of each edge corresponds to a 40° zenith angle.

resolution of 0.05 Hz. The sources in these skymaps are widely dispersed, filling most of a ±40° cone above the station. Indicated below each skymap are the local time of the measurement and the number of sources present simultaneously. With the radial velocity $v_{Rs} = \pi d_s/k = c d_s/2f$ given at a number of source points the plasma bulk velocity **v** can be calculated. A least-squares fitting procedure is applied to determine the three components of the velocity vector:

$$\text{Min}\ \varepsilon^2 = \sum_{s=1}^{S} \left[d_s - \frac{1}{\pi}\left(v_x k_{sx} + v_y k_{sy} + v_z k_{sz} \right) \right]^2 \tag{3}$$

where S is the total number of source points which must be 3 or larger to get a solution. The technique breaks down when there are less than three sources, when the velocities become very small, or when the sources are very tightly clustered. For routine drift analysis the individual skymaps are superimposed for 5 minute intervals (see Section 7) assuming that the bulk plasma velocity is constant during this time.

The observed drift for 10-11 March 1989 at Millstone Hill shows a consistent westward drift (Figure 13, lower panel) during nighttime until 0315 LT, when it starts to rotate clockwise (through north) making two full cycles within 4 hours. The horizontal velocity (middle panel) varies between 40 and 220 m/s, reaching its minimum in the early morning hours when the drift vector is rotating. About one hour earlier the vertical velocity v_Z (upper panel) reaches a maximum of 17 m/s downward. The error bars are generally small during nighttime when the ionosphere is rough and many reflection points exist (disturbed ionospheric conditions caused some bad measurements around local midnight). During daytime hours the error bars on the drift direction become large. This is typical for mid-latitude, where the ionospheric roughness (see Section 7) during the day is small, and the few reflection points are clustered in a relatively small area.

The situation is more favorable in the polar cap where the roughness is always sufficiently large to continuously sustain a large number of reflection points, as illustrated in Figures 14 and 15 for data from Qaanaaq, Greenland (87° N corrected geomagnetic) for summer and winter, respectively. The data for 25-27 June 1989, when the ionosphere was continuously sunlit, show very small error bars. The lower panel in Figure 14 shows again the drift direction as a function of time; the azimuth is measured from corrected geomagnetic north. For reference the antisunward direction is indicated, and it is evident that the plasma flow is essentially antisunward, with the exception of a few hours around 09 UT on 25 and 27 June. This is the typical signature of the polar cap two cell plasma convection for IMF B_Z south [Heppner and Maynard, 1987]. The measured horizontal flow velocities vary between 150 and 800 m/s, and the vertical velocities stay generally between ±40 m/s. A more irregular pattern exists during the three winter days in January 1989, displayed in Figure 15. There is clear sunward motion on January 4 from 03 to 23 UT which is generally ascribed to positive B_Z IMF conditions. The highest speeds are observed during periods of antisunward convection.

Several years of Digisonde drift observations at Qaanaaq and at Goose Bay, an auroral station at 65° N CG, have brought convincing evidence that this ground-based technique is monitoring the high latitude convection pattern [Buchau et al., 1988]. This has been further quantified by statistical analyses of the 1986, 87 and 88 Qaanaaq data in relation to the IMF components B_Z and B_Y [Cannon et al., 1989a, 1989b]. These studies show excellent agreement with the two cell model of Heppner and Maynard [1987]. For B_Z south and B_Y negative the average convection direction is rotated anticlockwise from the antisunward direction (Figure 16). For B_Z south and B_Y positive the average convection direction is rotated clockwise (Figure 17). The convection direction when B_Z is north has also been examined. When B_Z is north and B_Y negative the statistical significance of our results is low but the measurements support a severely distorted two cell convection model. For B_Z north and B_Y positive no well defined convection direction can be discerned. The velocity shears in the vicinity of sun aligned polar arcs that are frequently observed during this latter conditions [Carlson et al., 1984] could invalidate our analysis technique. To

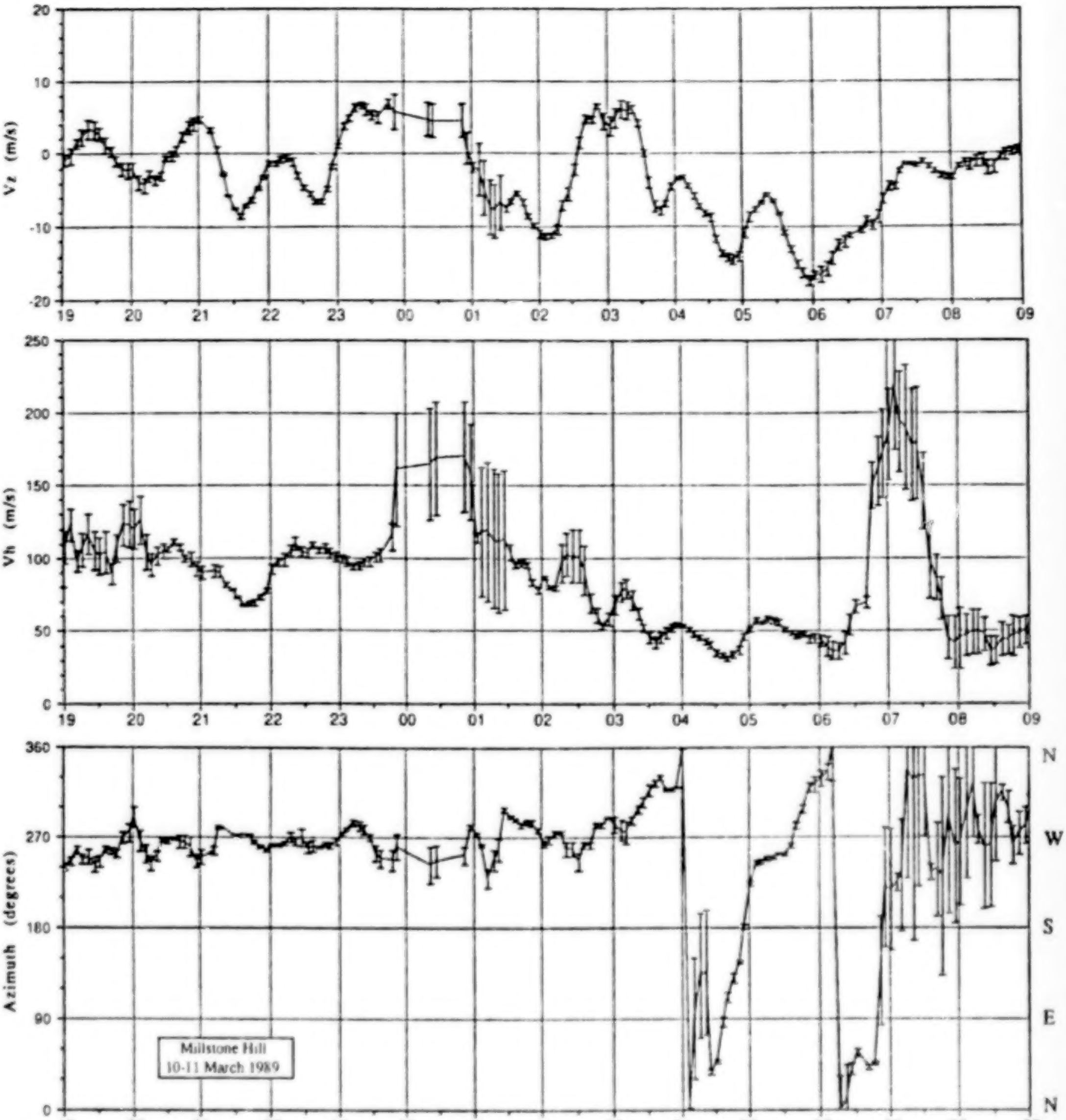


Figure 13 -- Drift velocity vector vertical (top), horizontal magnitude (middle) and azimuthal direction (bottom) components for Millstone Hill on 10-11 Mar 89. Velocity vectors are calculated every 5 minutes using five 20 second long measurements. Sparse data from 23:45 LT to 01:30 LT produce large uncertainties in the velocity. Sunrise occurs at 06:05 LT, after which the drift directions become uncertain.

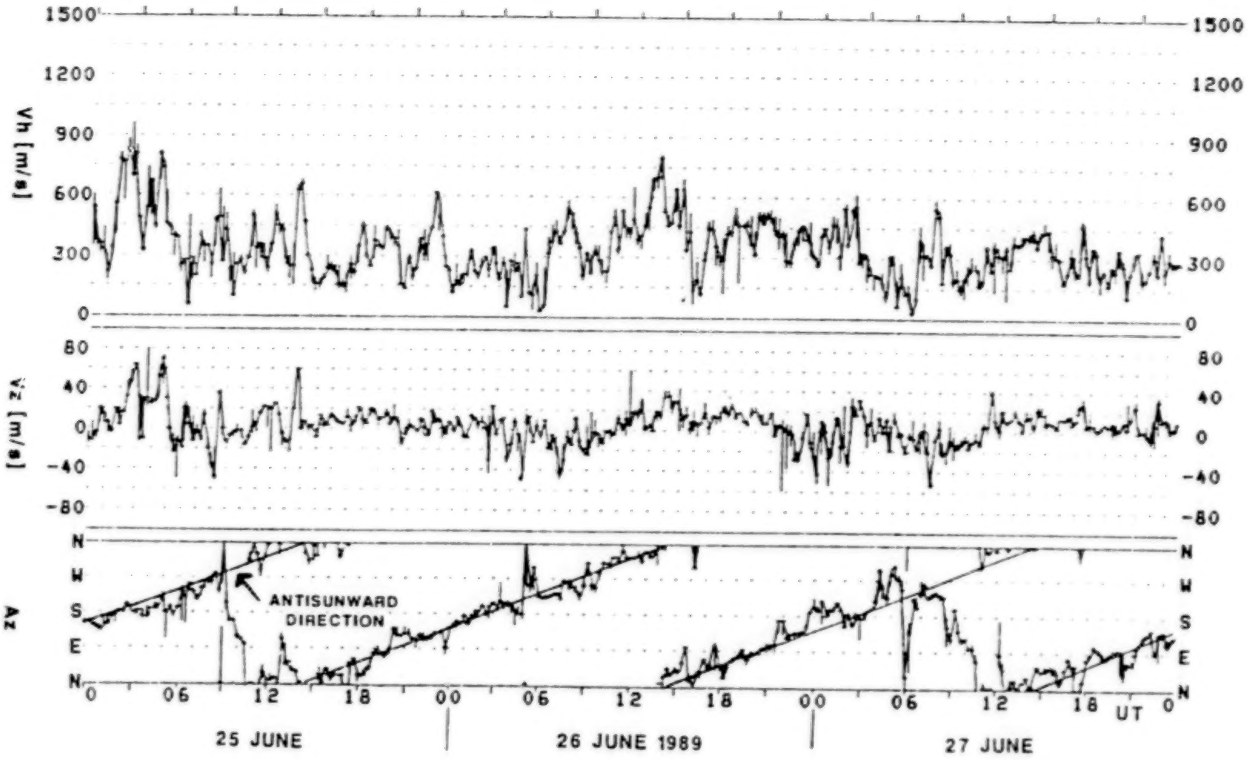


Figure 14 -- Time sequence of drift velocities measured by the Digisonde 256 at Qaanaaq, Greenland on 25-27 June 1989. Top, middle and bottom panels show the horizontal magnitude, the vertical component and the azimuthal direction respectively. Velocities are calculated every 15 minutes. The oblique lines in the bottom panel indicate the approximate antisunward direction. The plasma flow is essentially antisunward most of the time during these three days.

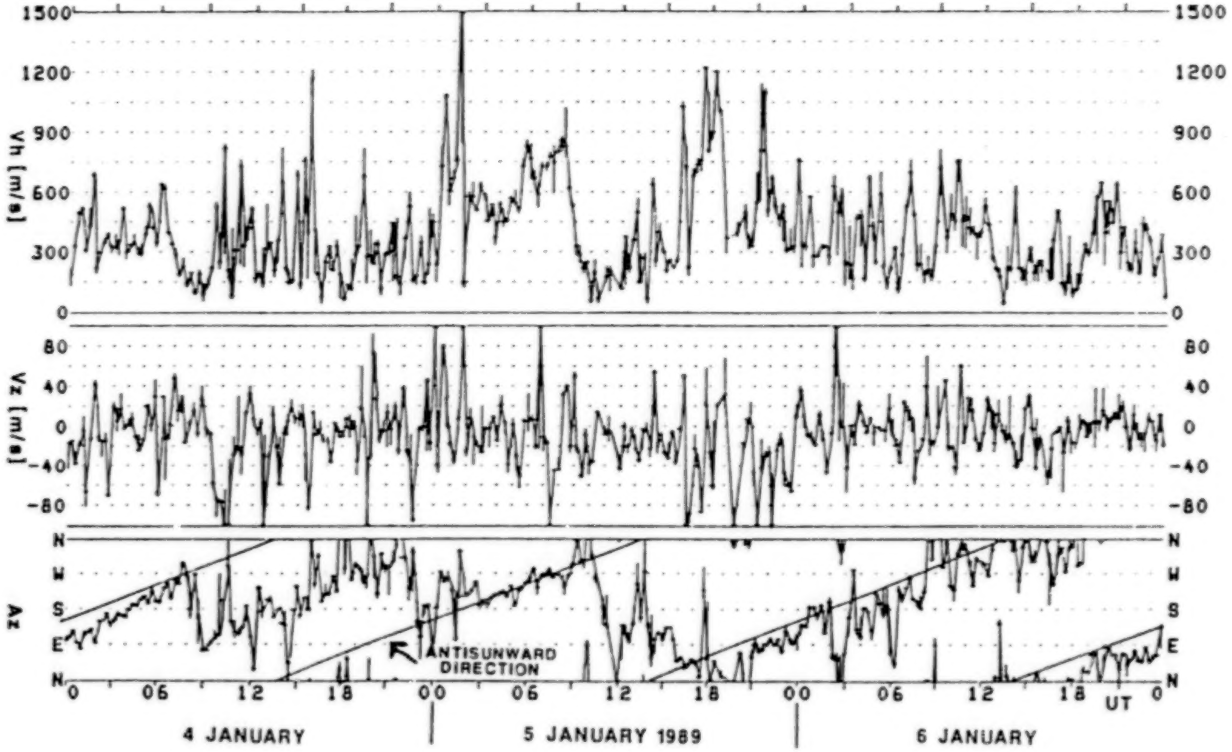


Figure 15 -- Time sequence of drift velocities measured by the Digisonde 256 at Qaanaaq, Greenland on 4-6 January 1989. Top, middle and bottom panels show the horizontal magnitude, the vertical component and the azimuthal direction respectively. Velocities are calculated every 15 minutes. The oblique lines in the bottom panel indicate the approximate antisunward direction. The plasma flow during these three days often diverges from the antisunward direction. Sunward flow predominates between 09 to 23 UT on 4 January.

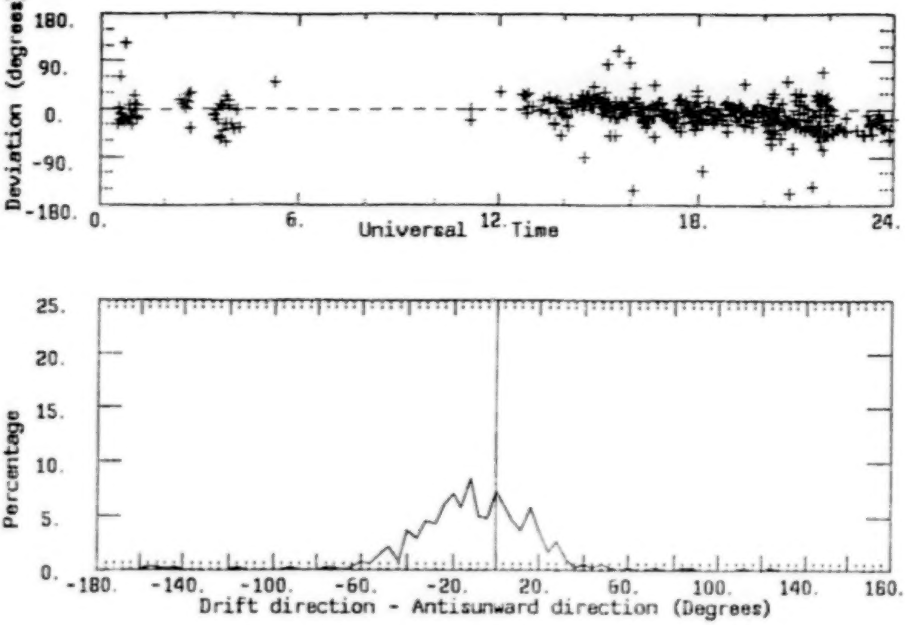


Figure 16 -- Deviation of the drift direction from the anti-sunward when Bz<0 and By < 0: top panel, temporal distribution of selected data, bottom panel, deviation of convection direction from the anti-sunward. The data was selected from 32 days of measurements in 1986,87 and 88 when the interplanetary magnetic field was stable .

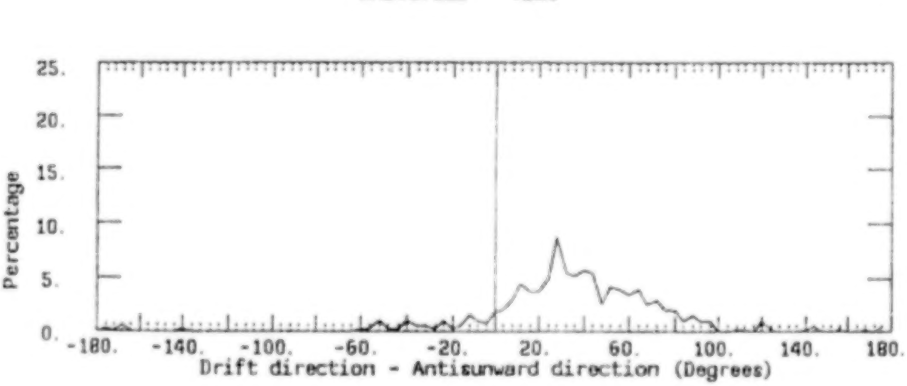


Figure 17 -- Deviation of the drift direction from the anti-sunward when Bz<0 and By > 0: top panel, temporal distribution of selected data, bottom panel, deviation of convection direction from the anti-sunward. The data was selected from 32 days of measurements in 1986,87 and 88 when the interplanetary magnetic field was stable .

properly analyze the Digisonde drift data during B_Z, B_Y positive conditions requires all-sky images of the subvisual arcs [Weber et al., 1984] that specify the location of the arcs.

At auroral latitudes, e.g. Goose Bay, the two cell convection pattern is mainly observed during the nighttime, when the clearly defined westward drift changes through south to an eastward drift at about local magnetic midnight. Of course what happens is that Goose Bay is rotating from the dusk cell through the Harang discontinuity to the dawn cell [Buchau et al., 1988]. The flow velocity reaches its minimum (upper panel in Figure 18) in the Harang discontinuity. It is interesting to note that the flow velocity increases rapidly from about 30 m/s to 300 m/s when Goose Bay moves into the trough. In the auroral oval the velocities decline rapidly reaching the values previously observed equatorward of the trough. The arrival times of the trough and the oval at Goose Bay are easily determined on the ionogram surveys discussed in Section 3.

6. GRAVITY WAVE STUDIES

Ionogram surveys (Figure 8) and electron density contour plots (Figure 9) are basic tools for the detection and identification of the traveling ionospheric disturbances associated with atmospheric gravity waves (AGW). The contour plots give the wave amplitude and period, but the wavelength and the tilt of the wave front cannot be obtained from the contour plot which presents the electron densities as function of time, not of horizontal distance. To determine the true wavelength, the velocity and the propagation direction of an AGW requires finding the instantaneous electron density surface over a horizontal extend comparable to its wavelength. The incoherent scatter radar could in principle make such measurements,but the acquisition and operating costs render these instruments unsuitable for global gravity wave studies and they also are too slow to do scanning.

The spaced-station technique (Figure 19) deploying several ionosondes separated by 50 km to a few hundred kilometers can determine the AGW parameters. In principle, clusters of analog ionosondes are quite adequate for this task, however, the logistics involved in synchronously running these many sounders and processing all the data are enormous. Also the presence of small and middle scale AGWs can introduce large errors.

Analysis of the Digisonde's drift mode data can lead to the calculation of the true wavelength and velocity of the AGW. We are currently trying to fit a set of sinusoidal surfaces through the observed source points for a given sounding frequency (i.e. plasma frequency) in an effort to determine the wavelengths and the orientations of these wave structures (Figure 20). The velocities can in principle be found from the Doppler frequencies of the source points, however the results may be flawed in the presence of plasma drift. More research is required to resolve this difficulty. Calculating the phases of the AGW, relative to the sounder zenith position, for different plasma densities allows one to determine the tilt of its wave front.

7. IONOSPHERIC TILT MEASUREMENTS

Oblique ionospheric HF propagation characteristics are relatively sensitive to horizontal gradients in ionization near the reflection region, since the resulting tilts affect the skip distance and the associated maximum usable frequency (MUF) [Munjal and Sales, 1987]. The Digisonde, operating in the drift mode, can measure the ionospheric tilt. Combining the tilt and vertical profile information improves the predictability of the optimum working frequency for a particular radio path, whether it be for communication or for over-the-horizon radar application. The new tilt technique described here also measures the irregularity structure or "roughness" of the ionosphere. Any theory dealing with the dynamic processes in the upper atmosphere should consider these measured variations in the ionospheric roughness.

Section 5 has discussed the skymaps generated from the Digisonde drift mode data. Each map locates the reflection points (sources) for a coherent dwell, typically 20 sec at mid-latitude. For the tilt and roughness analysis, all 20 sec skymaps for a 5 min observation period are superimposed resulting in a composite skymap. This sampling rate can resolve the effects of acoustic gravity

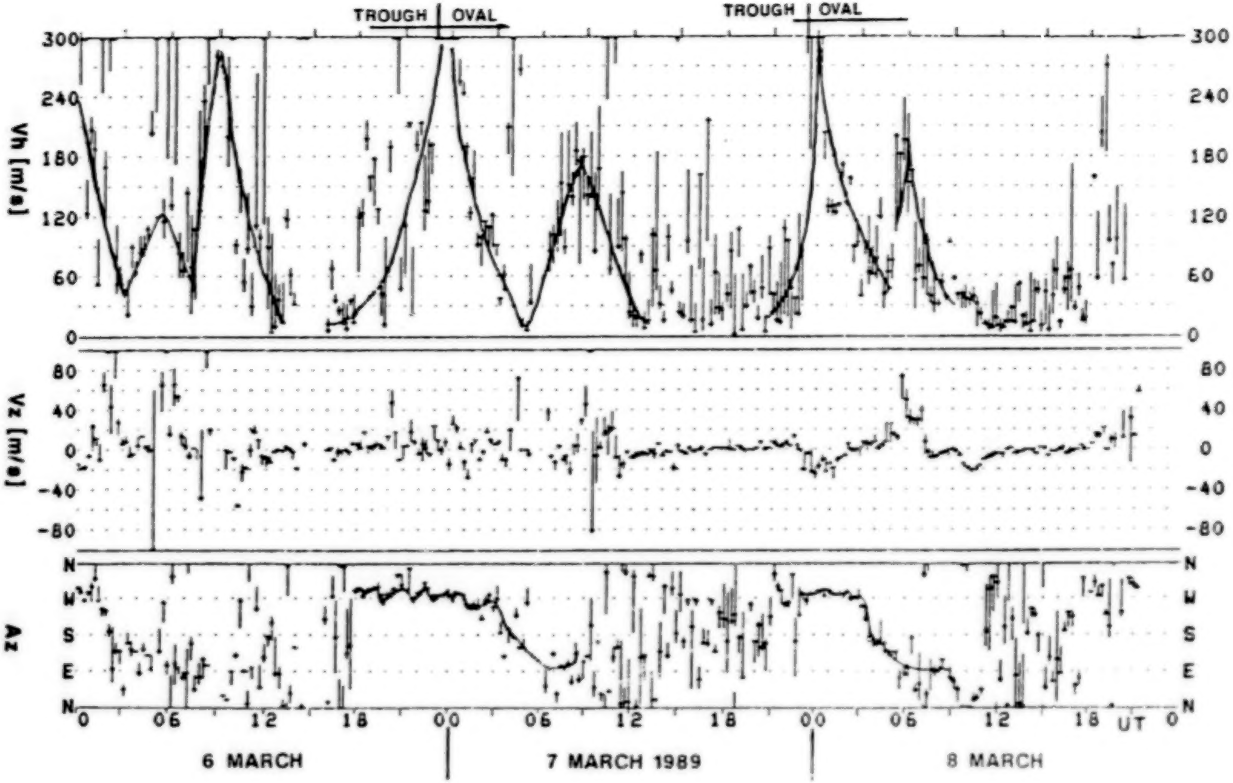


Figure 18 -- Time sequence of drift velocities measured by the Digisonde 256 at Goose Bay, Labrador on 6-8 March 1989. As Goose Bay enters the trough in the evening sector, the drift speed (top panel) increases until the auroral oval is entered. The drift direction (lower panel) is usually stable during the evening hours. The change from westward to eastward drift direction occurs close to local magnetic midnight (0330 UT) and coincides with very small drift magnitudes.

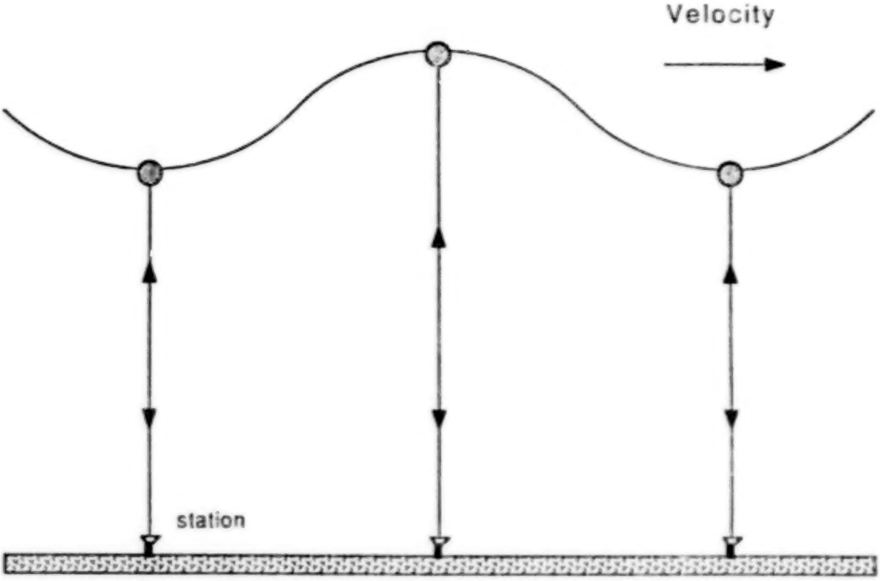


Figure 19 -- Spaced station technique for observing atmospheric gravity waves. Multiple stations simultaneously measure overhead electron density distributions. Comparison of data from coordinated stations can determine the amplitude, phase, wavelength and velocity of the gravity wave.

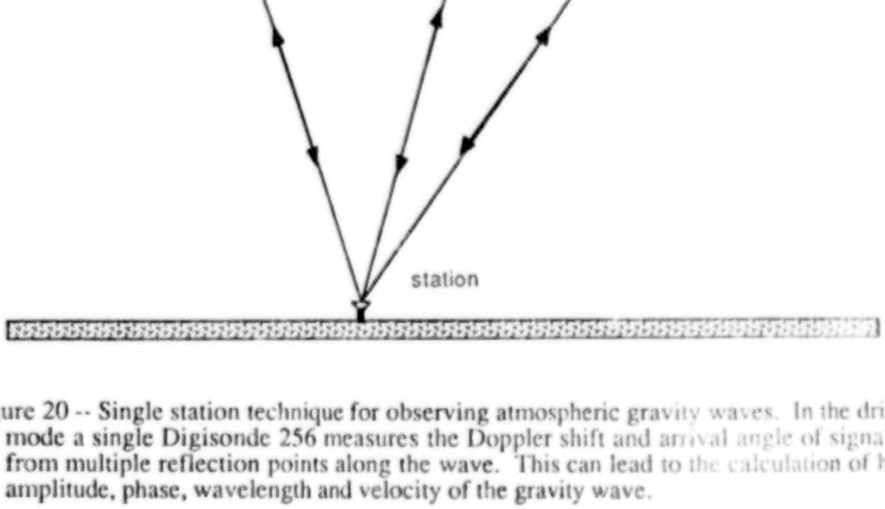


Figure 20 -- Single station technique for observing atmospheric gravity waves. In the drift mode a single Digisonde 256 measures the Doppler shift and arrival angle of signals from multiple reflection points along the wave. This can lead to the calculation of he amplitude, phase, wavelength and velocity of the gravity wave.

waves passing over the site. Figure 21 shows sample composite skymaps for Millstone Hill. The upper half illustrates the concentration of the sources in a very narrow angular region during daytime. During nighttime (lower half of Figure 21) the situation is quite different: the angular locations of the sources are widely spread over the "sky," often out to 30° for mid-latitude, and more in the auroral and polar cap regions.

These representative maps highlight the diurnal variation in the structure, or the roughness, of the reflecting layer. To quantify the spread of the sources we will later define a roughness index. The concept of a tilt of an isodensity surface is well defined during the daytime when all the echoes arrive within a narrow angular cone. The center of this cone, determined as the center of gravity (CG) of the skymap points, specifies the "average" reflection point marked as CG in Figure 21. The CG is displaced from the overhead position in the presence of a tilt. This displacement will occur even when the source distribution is very wide. The angle between the zenith and CG gives the tilt of the layer, and the angle between true north and the vertical plane through the sounder and CG defines the azimuthal orientation (bearing) of the tilt. Figure 22 is a data sample for three consecutive days showing the diurnal variations of the tilt angle and tilt bearing. These data are inherently noisy and the routine presentation of these data uses a one hour (12 point) running average to produce the smooth plots.

We expect to find a diurnal variation of the tilt in addition to any variations caused by AGWs passing over the site. This diurnal variation can be seen for some of the data in the above example. There are clear changes associated with the occurrence of sunset and sunrise in the ionosphere above the Digisonde site. The large cyclical variations in the tilt bearing during the night period were not expected and research is needed to understand the mechanism for these changes. The tilt angles vary from small tilts of of the order of one degree or less to a typical nighttime average of about 4°. These larger nighttime tilts may be related to the dominance of long period AGWs at these times.

Having located the CG of the distribution of reflecting sources we can now measure the spread of the sources about the CG point. We divide the skymap into eight equal angular sectors (45° each) about the CG as the origin. Within each sector a circular arc is drawn about the CG with radius such that one half of the points within the sector lie outside and the other half lie within this arc. These eight arcs define a curve which surrounds one half of all the sources on the particular skymap, and the measure of the solid angle in steradians (SR) of the area inside this curve is defined as the Roughness Index (RI). In order to avoid some peculiarities caused by a small number of points in a particular sector we have introduced a weighting factor proportional to the number of points in the particular sector.

The plot of the smoothed RIs for 3 consecutive days is presented in Figure 23. The RI varies from a value of less than 10 SR in the daytime to a considerably higher values at night, often reaching numbers greater than 100 SR. The daytime clustering of the reflection sources into a small angular area is best described as a smooth ionosphere with a low RI. In contrast, the night ionosphere with high RIs, corresponding to a spread distribution of sources, is characterized as rough. Using models of the irregularity structure makes it possible to determine the scale sizes that are likely to be associated with the changes in the ionosphere from smooth to rough. Our preliminary work indicates, as expected, a good correlation between the nighttime roughness increase and spread-F conditions. Future capabilities of the Digisonde will make it possible to simultaneously measure the tilt and roughness parameters at all heights from the E-region up to near the F-layer peak.

Acknowledgement

This research was in part supported by the Air Force Geophysics Laboratory. The authors thank J. Buchau for his cooperation and numerous informative discussions. The Digisonde data at Goose Bay were collected by Canadian Marconi (B. Gill), and at Qaanaaq by the Danish Meteorological Institute (S. E. Ascanius). We thank R. R. Gamache, C. G. Dozois, P. S. Cannon and G. Crowley for their assistance in preparing this paper.

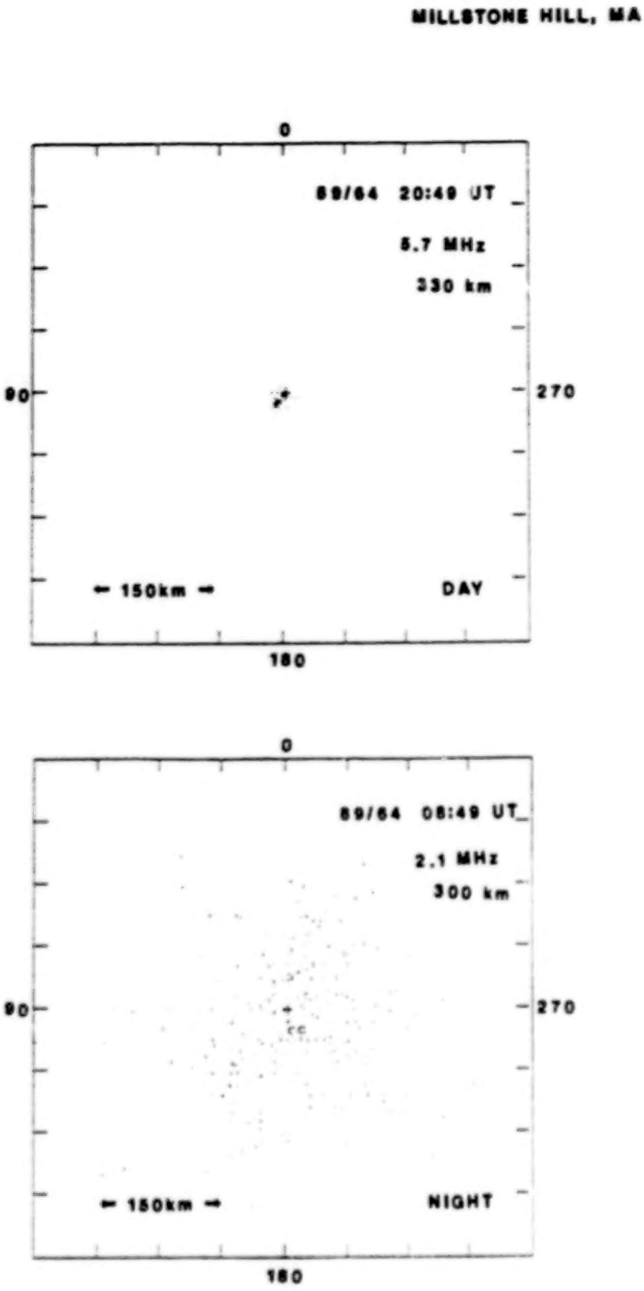


Figure 21 -- Sample skymaps from the ULCAR Digisonde site at Millstone Hill, Westford, MA. The upper panel shows a typical daytime map where all the sources fall within a very small reflecting area. The lower panel is a typical nighttime map showing the increased spread in the angular location of the reflection sources. The center of gravity of all the source points are indicated with the "CG" symbol on each map.

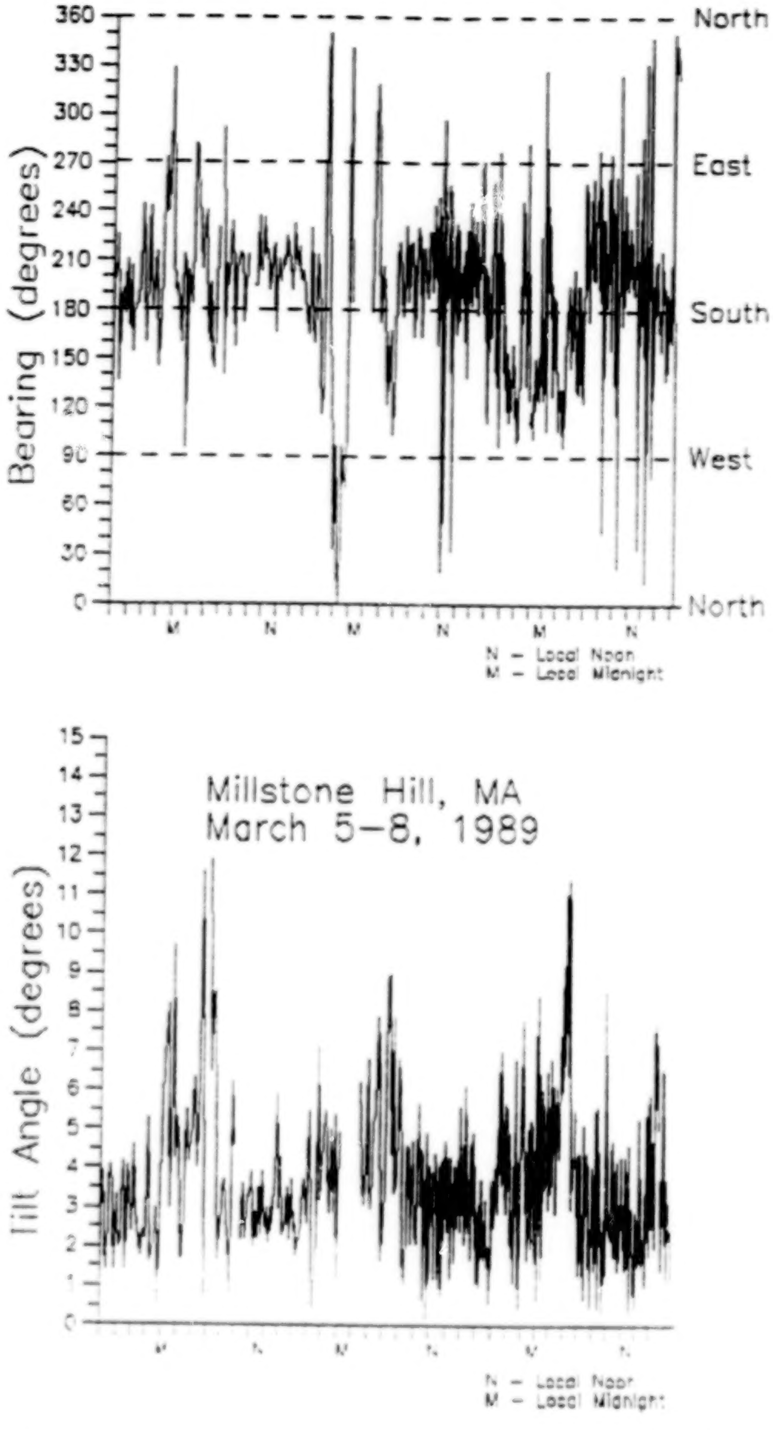


Figure 22 -- Data sample from the ULCAR Digisonde site at Millstone Hill showing the diurnal variation of the tilt angle and bearing for a three day period of data with no smoothing.

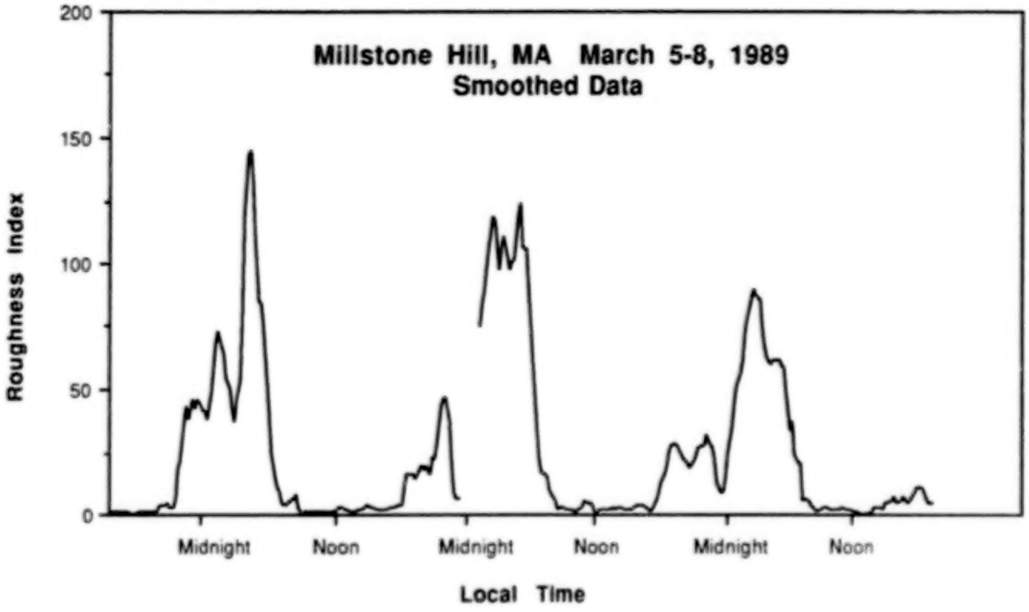


Figure 23 -- The smoothed variations of the roughness index (RI) is shown for the same three days as above. Each day shows a strong diurnal variation of the RI. The roughness index measures the solid angle for the area of the reflecting sources for each five minute period.

REFERENCES

Bibl, K., W. Pfister, B. W. Reinisch and G. S. Sales, 1975, "Velocities of Small and Medium Scale Ionospheric Irregularities Deduced from Doppler and Arrival Angle Measurements," COSPAR, Space Research XV, pp. 405-411.

Bibl, K. and B. W. Reinisch, 1978, "The Universal Digital Ionosonde," Radio Science, 13, No. 3, pp. 519-530

Bibl, K., B. W. Reinisch and D. F. Kitrosser, 1981, "Digisonde 256 - General Description of the Compact Digital Ionospheric Sounder," First Edition, University of Lowell Center for Atmospheric Research.

Breit, G. and M. A. Tuve, 1926, "A test of the existence of the conducting layer," Phys. Rev., 28, 554.

Buchau J., D. N. Anderson, E. J. Weber, B. W. Reinisch and C. G. Dozois, 1988, "Polar Cap Plasma Convection Measurements and Their Relevance to the Real Time Modeling of the High Latitude Ionosphere,".Radio.Science, 23, No. 4, 521.

Buchau, J., B. W. Reinisch, E. J. Weber and J. G. Moore, 1983, "Structure and dynamics of the winter polar cap F region," Radio Science, 18, No. 6, pp. 995-1010.

Buchau, J., 1989, private communication.

Buonsanto, M., 1989, private communication.

Cannon P. S., B. W. Reinisch, J. Buchau, T. Bullett, and C. Dozois, 1989a, 'Digital Ionosonde Measurements of High Latitude Ionospheric Convection," Proceedings of IAGA Scientific Assembly, July 1989.

Cannon P. S., B. W. Reinisch, J. Buchau and T. Bullett, 1989b, "A Statistical Validation of the Digisonde Technique for Polar Cap F-Region Convection Measurements.", Submitted to J. Geophys. Res. 1989.

Carlson, H. C., Jr., V. B. Wickwar, E. J. Weber, J. Buchau, J. G. Moore and W. Whiting, 1984, "Plasma characteristics of polar cap F layer arcs," Geophys. Res. Lett., 11, p. 895.

Heppner, J. P. and N. C. Maynard, 1987, "Empirical High-Latitude Electric Field Models, "J Geophys. Res., 92, A5, 4467-4489.

Huang, Xueqin and B. W. Reinisch, 1982, "Automatic Calculation of Electron Density Profiles from Digital Ionograms. 1. Automatic O and X Trace Identification for Topside Ionograms," Radio Science, 17, No. 2, pp. 421-434, March-April 1982.

Huang, Xueqin, 1989, "A Layer Model of the Profile-Fitting Method for Real Time Reduction of Digital Ionograms" (to be published).

Kuklinski, W. S., K. Chandra and B. W. Reinisch, 1988, "Status of DORIS Automatic Scaling, Interim Technical Report, NorthWest Research Associates, Inc., Bellevue, WA.

Munjal, M. and G. S. Sales, 1987, "Improved OTH-B Radar Targeting Using Real-Time Ionospheric Profile Measurements," RADC Technical Report (to be published)

Piggott, W. R. and K. Rawer, 1972, U.R.S.I. Handbook of Ionogram Interpretation and Reduction, Rep. UAG-23, 2nd ed., World Data Center A, Natl. Oceanic and Atmos. Admin., Boulder, CO.

Reinisch, B. W., 1986, "New Techniques in Ground-Based Ionospheric Sounding Studies," Radio Science, 21, No. 3, pp. 331-341.

Reinisch, B. W., K. Bibl, M. Ahmed, H. Soicher, F. Gorman and J. C. Jodogne, 1984, "Multipath and Doppler Observations During Transatlantic Digital HF Propagation Experiments," AGARD Conf. Proceedings, 363, 12-1 to 12-11.

Reinisch, B. W., J. Buchau and E. J. Weber, 1987, "Digital Ionosonde Observations of the Polar Cap F Region Convection," Physica Scripta, 36, pp. 372-377.

Reinisch, B. W., R. R. Gamache and L. G. Bossy, 1989, "Ionospheric Characteristics for IRI in Real Time," Adv. in Space Res., in print.

Reinisch, B. W., R. R. Gamache, X. Huang and L. F. McNamara, 1988, "Real Time Electron Density Profiles from Ionograms," Adv. in Space Res. 8, 4, pp. (4)63-(4)72.

Reinisch, B. W. and Xueqin Huang, 1983, "Automatic Calculation of Electron Density Profiles from Digital Ionograms. 2. True Height Inversion of Topside Ionograms with the Profile-Fitting Method," Radio Science, 18, No. 3, pp. 477-492, May-June 1983.

Snyder, M. A., 1966, Chevyshev Methods in Numerical Approximation, Prentice-Hall, Englewood Cliffs, NJ.

Tang, J., R. R. Gamache and B. W. Reinisch, 1988, "Progress on ARTIST Improvements," Interim Technical Report, NorthWest Research Associates, Inc., Bellevue, WA.

Titheridge, J. E., 1985, "Ionogram Analysis with the Generalized Program POLAN," World Data Center A for Solar Terrestrial Physics, Rep. UAG-93.

Weber, E. J., J. Buchau, J. G. Moore, J. R. Sharber, R. C. Livingston, J. D. Winningham and B. W. Reinisch, 1984, "F-layer ionization patches in the polar cap," J. Geophys. Res. 89, 1683.

Xu, Ling, B. W. Reinisch and M. Buonsanto, 1989, "Ionosonde and Incoherent Scatter Radar Measurements of Vertical Electron Density Profiles," Development of IRI-90 URSI/COSPAR Workshop, Abingdon, England, 7-9 August 1989.

Chapter 14

The HF Doppler technique for monitoring Transient Ionospheric Disturbances.

Introduction.

Many types of transient disturbances in the ionosphere, such as those resulting from solar flare activity or the passage of travelling disturbances (TIDs), distort the iso-ionic contours and so produce variations in the reflection height of an HF signal. This change in reflection height gives rise to a change in the phase path, consequently a Doppler frequency shift will be induced into the reflected signal. The magnitude of the frequency shift will be proportional to the rate of change of the phase path. By monitoring the frequency of the reflected signal, the presence of a disturbance which produces height changes can be detected and its time history recorded. If measurements are made on three or more propagation paths whose reflection points are suitably spaced, then a travelling disturbance will produce Doppler signatures which are time displaced on the signals received over the three propagation paths. If no dispersion occurs as the disturbance propagates, the horizontal component of its velocity can be determined from the three time delays and a knowledge of the reflection point separation by simple triangulation.

The Doppler technique provides a low cost, effective method of monitoring transient changes in the ionosphere. It does not always yield an accurate measure of the TID velocity because the driving internal acoustic gravity wave is generally dispersive. Nevertheless, the technique can produce valuable information as indicated later in this chapter.

Choice of operating frequency.

An HF radiowave will be reflected from the ionosphere provided its frequency is less than the F-layer critical frequency. In the Doppler technique, the receiver and transmitter are usually separated by distances of not more than 200 km and so propagation is at steep (almost vertical) incidence. The magnitude of the Doppler shift Δf is related to the rate of change of phase path P by

$$\Delta f = -\frac{1}{\lambda}\frac{dP}{dt}$$

where λ is the wavelength.

A change in phase path can occur either due to a change in the reflection height or because of a change in the refractive index along the ray path. It has been shown (Davies 1965, 1968) that for changes in reflection height the magnitude of Δf is directly proportional to the wave frequency, thus

$$\Delta f \propto f$$

If the frequency shift is induced by refractive index changes then

$$\Delta f \propto \frac{1}{f}$$

For reflections from the E and F layers of the ionosphere, Δf is found to be proportional to f indicating that changes in the reflection height are dominant in producing the Doppler shift.

Since the magnitude of the Doppler shift is proportional to the wave frequency, it would appear that the operating frequency should be chosen as close to the critical frequency as possible. This is not the case, since this choice results in the wave penetrating the layer as the diurnal changes in critical frequency occur. However, since the amplitude of all wave-like disturbances increase with increasing height, ie. decreasing atmospheric density, the displacement of the reflection level and hence the Doppler shift, increase with height, ie. with sounding wave frequency. For radio waves reflected from the mid-latitude F-region at steep incidence, a typical medium-scale TID will produce Doppler shifts of between -0.5 to +0.5Hz. A spectral resolution of 0.01Hz is more than adequate for studies of such events and a resolution of 0.05Hz is normally employed. If the measuring system is designed to measure frequency shifts of the order of 0.01Hz, it is not necessary to operate very close to the F-region critical frequencies and it is best to work at least 1 or 2 MHz below the critical frequency. The frequency should be selected by reference to an ionogram and chosen so that reflection occurs well away from regions of large group retardation. Since the critical frequency differs markedly during the day and night, it is not normally possible to operate a Doppler system for 24 hours on only a single frequency. Two frequencies are usually required, one for night and one for day-time conditions. It is rather difficult to measure the Doppler signature of E-region reflections since the Doppler shifts are very small due to the small displacements of the reflection height and the lower frequencies necessary to obtain E-region reflections. It should be noted that the sporadic E-layer produces a very stable Doppler signature with little or no variations over considerable periods of time. For general studies of TIDs, F-region reflections should be employed.

The Experimental System.

a) Transmitter.

The transmitter consists of any suitable RF amplifier which is capable of producing a CW signal of about 20 watts of RF power at the required HF frequency. Some amateur radio transmitters are quite suitable for this purpose. The transmitter drive is generated from a frequency synthesizer capable of producing the required frequencies and stable to about 1 part in 10^9. A good quality crystal oscillator can achieve this stability and the internal crystal oscillators of most modern synthesizers meet this stability requirement. For example, if we are to measure shifts of say 0.01Hz in a 5MHz signal a measurement accuracy of 1 part in 5 10^8 is needed.

The transmitting antenna normally consists of a simple half wavelength horizontal dipole suspended a quarter of a wavelength above the ground. A lower height can be employed, the requirement being that there is a major radiation lobe in the vertical direction. This ensures that power is radiated at an angle suitable for the steep incidence path between the transmitter and receiver. The dipole antenna is aligned so that it is end on to the direction of the receiver as this reduces the radiation in that direction and so reduces the ground wave field strength at the receiver. If continuous operation during day and night is required, provision should be made for radiating two frequencies to allow for the changes in critical frequency. For two frequencies it is usually easier, cheaper and more reliable to duplicate a complete transmitter installation than to introduce an automatic frequency changer into the system.

b) Receiver.

A standard HF communications receiver can be employed, provided the local oscillator is sufficiently stable. In general, this will not be the case and arrangements will have to be made to generate this frequency from a synthesizer of similar stability to that employed at the transmitter. When the synthesized frequency is set equal to the transmitted frequency minus the local oscillator frequency, the nominal IF frequency will be obtained from the IF amplifier. This frequency is down converted to base band by mixing with a synthesized IF frequency (see fig 1). This mixing is done in phase and in quadrature with two mixers as indicated in the figure. The real and imaginary channels are fed to two A/D converters which digitise the base band frequency at a rate of between 30 to 50 samples per second. These digitized samples are recorded onto magnetic tape or disc for further processing.

Although amplitude information is not required for a simple Doppler experiment, some form of AGC must be applied to the receiver to overcome the large changes in signal amplitude due to fading. The normal receiver AGC system is generally adequate but an attenuator controlled by the computer can be incorporated. The attenuator is changed in such a way as to keep the receiver output amplitude approximately constant during the course of the measurements.

The frequency of the signal is obtained by applying a FFT to the digitized data. This is done off-line although some capability for viewing the data on site is generally desirable. Samples of the data, representing 20 seconds of observing time are Fourier transformed. This provides a frequency resolution of 0.05 Hz within a spectral range of 30Hz. Clearly the amount of

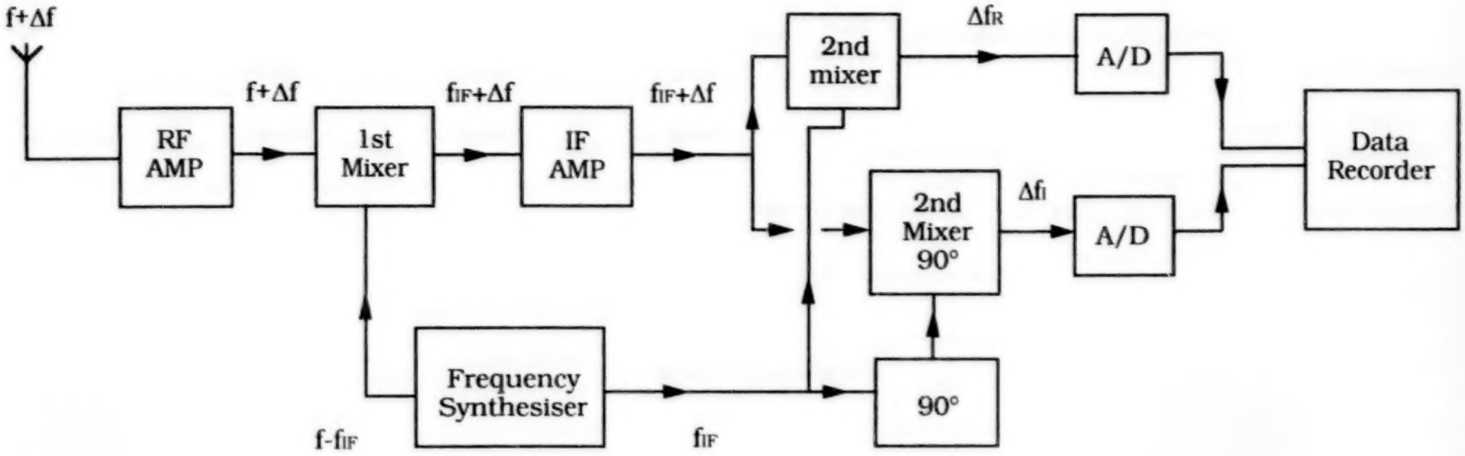


Fig. 1 Block diagram of Receiver elements

data included in each transform can be varied and so the frequency resolution can be changed. A sample of the analysed data is reproduced in fig.2, which contains examples of medium scale TIDs. When 3 or more transmitters are employed for measurements of TID velocities, it is usual to off-set the frequency of each transmitter from its nominal value by a few Hertz. Each transmitter is assigned its own unique off-set so that it can be easily recognized in the subsequent Fourier analysis. Four or five transmission off-sets can be accommodated within a 15 Hz bandwidth so they are within the passband of the receiver even at its narrowest setting. The receiver bandwidth should be as narrow as possible, say 100 to 200 Hz, to reduce unwanted signals and noise.

The receiving antenna can be of any design provided it is sensitive to signals arriving at steep angles of incidence. A convenient antenna is a short active dipole mounted horizontally some 5 to 10 meters above the ground. This produces the required polar diagram and responds well to any frequency within the HF band without adjustment.

The transmitter and receiver sites are usually separated by about 60 to 100 km to reduce the ground wave field strength at the receiver. If TID velocities are to be measured, then the spacing must be large enough to produce significant time delays between the signatures of the event on the three or more propagation paths employed. For medium scale TIDs the horizontal component of the velocity is about 300 ms^{-1} and for a reflection point separation of 30km this produces a maximum time delay of 100 sec. For large scale TIDs with periods of approximately 1 hr, the horizontal velocity can be as high as 1000ms^{-1} For studies of these disturbances, larger spacings of not less than 60 km between reflection points should be adopted. This corresponds to a maximum time delay of about 1 min. Since the TIDs are dispersive, the accuracy of the velocity measurement depends critically on how well the Doppler signatures on the various propagation paths can be correlated.

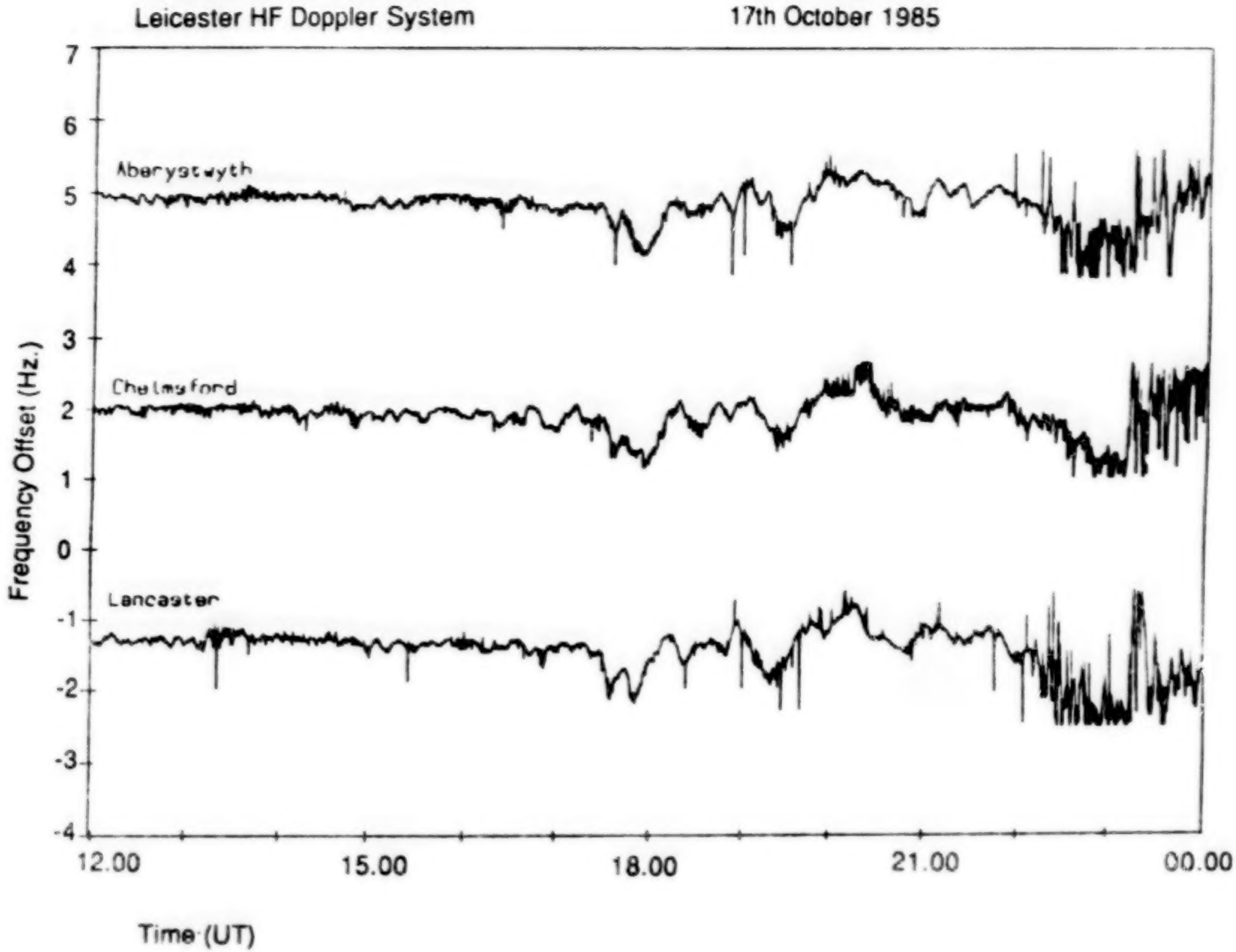


Fig 2. Example of three station Doppler variations from digital system.
Note wave activity from 18.00 to 21.00 UT.

389

c) Control computer and data logger.

Almost any small computer can be adapted as a controller and data logger for the receiver system. The following are the main considerations when choosing and/or designing a suitable computer:-

i) Interface with the receiver. This is not essential unless the receiver is to be switched between various operating frequencies on an automatic basis (eg. day/night usually have different requirements). On modern receivers this will often be an RS232 interface which can be connected to a serial communications port of the computer. Some receivers may require parallel control interfacing.

ii) A to D conversion. Two analogue to digital converters are required which can be simultaneously sampled at a fixed sample rate (approx. 30-50 samples per second is usual). The conversion time must be such that the input signals do not change substantially during the conversion. With modern converter ICs this is not a problem, since very fast (tens of μs per conversion) conversions can be achieved at little cost.

iii). Data Storage. Large quantities of data can be generated by the receiver system. For example, at 50 samples per second with 8 byte ADCs, and assuming continuous operation, 100 bytes/second $\equiv$ 360 kbytes/hour $\equiv$ 8.64 Mbytes/day are produced. Large storage capacity is therefore needed. If a hard disc is employed (an IBM PC/AT typically has a 30 Mbyte disc), this will need to be periodically off-loaded onto some other medium. Floppy discs are impractical in view of the number which would be required for this amount of data. Cartridge streamer tapes with capacities in the region of 60-150 Mbytes are readily available which fulfill this need. However, if the data is to be transferred to another (large) machine for subsequent processing and analysis then it may be advantageous to employ a conventional $^1/_2$" tape drive (20-30 Mbytes per reel).

iv) Analysis. FFT analysis can be performed on the data collecting/control computer if necessary. In view of the processing times required in such small machines, it is normally preferable to transfer the data to a mainframe or mini computer for subsequent analysis.

The requirements outlined above can be fulfilled by many small desk top computer systems, eg. IBM PC or compatibles. However, a simple microprocessor system is quite adequate and cost effective. A block diagram of one such system, as used at Leicester University, UK, is shown in Fig.3.

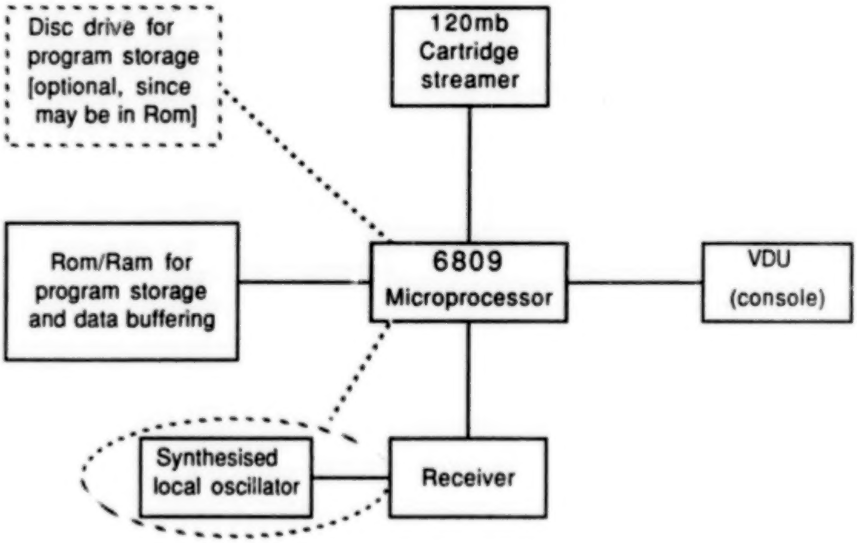

Disc drive for program storage [optional, since may be in Rom]
120mb Cartridge streamer
Rom/Ram for program storage and data buffering
6809 Microprocessor
VDU (console)
Synthesised local oscillator
Receiver

Fig. 3 Block diagram of Receiver control and data loging system.

Data Analysis

a) Determination of TID velocities.

If a TID gives rise to well correlated signatures on 3 or more propagation paths and the locations of the reflection points are known, the horizontal speed and direction of the disturbance across the measuring array can be determined as follows:

Let t_a, t_b and t_c be the relative times at which a characteristic wave of phase φ_o was observed on the frequency-time recordings for three spaced propagation paths. Let the reflection points be at $\underline{r}_a$, $\underline{r}_b$ and $\underline{r}_c$. It is usual to assume that the reflection points are vertically above the mid-points of the transmission paths and that they lie in the same horizontal plane. If the disturbance approximates to an infinite plane wave of the form $\exp[i(\underline{k}.\underline{r}- \omega t)]$ and that it does not change its shape as it propagates, then its characteristic phase φ_o can be written as

$$\varphi_o = \underline{k}.\underline{r}_a - \omega t_a$$
$$\varphi_o = \underline{k}.\underline{r}_b - \omega t_b$$
$$\varphi_o = \underline{k}.\underline{r}_c - \omega t_c$$

Defining $t_o = \dfrac{\varphi_o}{\omega}$ and $\underline{a} = \dfrac{\underline{k}}{\omega}$ these equations become

$$t_o = \underline{a}.\underline{r}_a - t_a$$
$$t_o = \underline{a}.\underline{r}_b - t_b$$
$$t_o = \underline{a}.\underline{r}_c - t_c$$

Since $\underline{k}$ is horizontal, $\underline{a}$ is horizontal and there are only 3 unknowns (a_x, a_y and t_o) in the equations. Solving for a_x and a_y

$$a_x = \frac{y_{ac}\, t_{bc} - y_{bc}\, t_{ac}}{y_{ac}\, x_{bc} - y_{bc}\, x_{ac}} \qquad \text{where} \quad x_{ac} = x_a - x_c, \text{ etc}$$

$$a_y = \frac{x_{cb}\, t_{ac} - x_{ac}\, t_{bc}}{y_{ac}\, x_{bc} - y_{bc}\, x_{ac}}$$

Since $\underline{a}.\underline{v} = 1$ and $\underline{a}$ is parallel to $\underline{v}$

$$v_x = \frac{a_x}{a^2} \quad \text{and} \quad v_y = \frac{a_y}{a^2}$$

The horizontal trace speed is then given by

$$v_h = \frac{1}{|\underline{a}_h|}$$

and the azimuth angle of propagation Θ by

$$\Theta = \mathrm{Tan}^{-1}\left(\frac{a_x}{a_y}\right)$$

Since the delay times are measured and the location of the reflection points are known, v_h and Θ can be determined.

It is questionable whether the phase or group trace velocity is measured by the Doppler experiment (Hines 1974) and many authors avoid the problem by simply referring to a 'horizontal velocity'. Experimental configurations with reflection points separated by tens of kilometres derive velocities which are close to the phase trace velocity, but this is not quite the same as observing the progression of a single crest or trough of one frequency component. Experiments in which the observation points are separated by several hundred kilometres obtain velocities close to the group trace velocity.

b) Spectral Analysis.

The time delays between the various Doppler traces have to be determined and several techniques are available for this task. The simplest method is to produce hardcopy of the Doppler-time variations for each propagation path (similar to that reproduced in Fig.2) and to identify features which appear on all three (or more) by eye. The time delays can then be determined directly from the recordings. This very simple procedure can be adopted only when there are very well defined features with large relative time delays and even then, high accuracy is difficult to achieve. It is also possible to calculate the cross correlation coefficient between the variations on the various propagation paths but again it is difficult to obtain an accurate result, particularly if the disturbances are dispersive which will generally be the case. The method commonly adopted for determining the time delays is that of cross spectral analysis in which a given interval of data is Fourier analysed to obtain the major frequency components present. The phase differences between the same frequency component in the data for each propagation path is calculated. From this the time differences for each path and for each spectral component are obtained.

c) Cross Spectral Analysis.

Cross spectral techniques have been applied to measurements of ionospheric motions with spaced receivers by several authors [see for example, Jones and Maude (1968, 1972)]. The cross spectrum of two signals $x_1(t)$ and $x_2(t)$ is defined as

$$S_{12}(f) = x_1^*(f).\, x_2(f)$$

where $x_1(f)$ and $x_2(f)$ are Fourier transforms of $x_1(t)$ and $x_2(t)$, and * denotes the complex conjugate. $S_{12}(f)$ is a complex quantity whose real and imaginary parts are called the cospectrum and quadrature spectrum

$$S_{12}(f) = L_{12}(f) - iQ_{12}(f)$$

This may also be expressed as the product of amplitude and phase functions called the cross-amplitude and phase-spectra.

$$S_{12}(f) = A_{12}(f) \quad e^{iF_{12}(f)}$$

where $\quad A_{12}^2(f) = L_{12}^2(f) + Q_{12}^2(f)$

and $\quad F_{12}(f) = \arctan [\, -Q_{12}(f) \,/\, L_{12}(f) \,]$

The cross-amplitude spectrum indicates whether frequency components in one time series are associated with large or small amplitudes at the same frequency in the other series. The phase spectrum indicates the phase relationship between the frequency components of the two time series. The phase differences can be converted into time delays. The velocities of disturbances are determined from these time delays and the geometry of the reflection points as described in the previous section.

The squared coherency spectrum, referred to as the "coherency" can be written as

$$K_{12}^2(f) = \frac{L_{12}^2(f) + Q_{12}^2(f)}{|x_1(f)|^2 \; |x_2(f)|^2}$$

The coherency $K_{12}(f)$ provides a measure of the noise in the two spectra at a given frequency. When there is perfect correlation the noise is zero and the coherency is unity. Similarly, when the two signals are uncorrelated at a given frequency their coherency is zero. The coherency can, therefore, provide an indication of the confidence that can be placed on estimates of the phase spectrum and hence, on the corresponding velocity estimates.

The Fourier analysis is limited in that the variance of the spectral estimates is not decreased by employing longer data sets. Smoothing the estimates in the frequency domain will decrease their variance and produce a more reliable estimate of the spectrum. The discrete nature of the data, due to the digital recording and subsequent Fourier analysis, makes it impossible to recognise harmonic components whose frequencies are not integer multiples of the frequency spacing. The power at such a component is therefore distributed over a range of frequencies. This is known as leakage. The data are analysed in discrete time blocks and these can be effectively operated upon by rectangular windows, since the Fourier transform of a rectangular window is the sinc function. The large side lobes of this function decay slowly, so leakage may be appreciable. Leakage may be suppressed, however, by smoothing in the time or frequency domains. This corresponds to the use of a convergence factor. Sloan (1969) and Yuen (1977) favour quadratic smoothing (ie. spectral smoothing) against linear (ie. data window) smoothing. Both systems reduce leakage, but in the former case the loss of resolution from smoothing is offset by a reduction in variance.

A Hanning window, which incorporates weights of (0.25, 0.5, 0.25) has been found to be effective for smoothing all spectral estimates in the frequency domain. The width of the main lobe of this window is $W_{main} = 4\Delta f$ where Δf is the frequency spacing of the spectral estimates. Its equivalent bandwidth is (Blackman and Tukey, 1958; Jenkins and Watts, 1968) given by

$$W_E = 0.67\ W_{main} = 2.67\ \Delta f$$

Blackman and Tukey (1958) draw attention to the need for rejection of zero frequency components (including linear trends) from the data before spectral analysis. Even small displacements of the mean from zero can contribute large peaks at zero frequency, and at neighbouring frequencies due to leakage. This is a particularly important criterion for cross spectral analysis where tests indicated that the phase spectrum was particularly sensitive to leakage from zero frequency components.

The data can have near-zero phase spectra, in which case a simple replacement of the sample co-spectrum and quadrature spectrum by their smoothed counterparts in the Equations for $A_{12}^2(f)$ and $F_{12}(f)$ produces the most accurate cross spectral results (Tick, 1967). Thus

$$\bar{F}_{12}(f) = \arctan\left[-\bar{Q}_{12}(f) \;/\; \bar{L}_{12}(f)\right]$$

$$\bar{K}_{12}^{2}(f) = \frac{(\bar{L}_{12}(f))^2 + (\bar{Q}_{12}(f))^2}{|x_1(f)|^2 \quad |x_2(f)|^2}$$

The coherency $K_{12}(f)$ may be transformed to the variable $Y_{12}(f)$ = arctanh $(K_{12}(f))$ via the Fisher transformation, then the 95% confidence interval for $Y_{12}(f)$ can be determined which is independent of frequency (Jenkins and Watts, 1968). Approximately 95% confidence intervals ($\pm\Delta\varphi$) for the smoothed phase spectrum with varying degrees of freedom (v) can be calculated from

$$\tan\bar{F}_{12} \pm 1.96\,\frac{\sec^4\bar{F}_{12}}{v}\left(\frac{1}{\bar{K}_{12}^2} - 1\right)$$

The confidence interval is parametric in $\bar{K}_{12}^2$ but approximately independent of $\bar{F}_{12}$ (Jenkins and Watts, 1968).

d) Rejection Criteria for Phase and Velocity Estimates.

3-station Doppler data provide 3 independent phase spectral estimates, any two of which give a measure of the horizontal velocity of the disturbance at a given frequency. The three velocity estimates thus obtained were rejected by Jones and Maude (1972) if their azimuths differed by more than 10 degrees. The coherency spectrum provides another measure of noise, but Jones and Maude (1972) found that coherency criteria rejected different (fewer) velocity estimates. They suggest that this is simply due to the unreliability of coherency estimates when they are low.

The azimuthal effect of small timing errors is largest in situations where the timing errors and time delays between Doppler traces are comparable. When cross spectral methods are applied to disturbances travelling at high speeds, the 10 degree criterion would thus tend to reject much valid data.

Coherency criteria are therefore more appropriate for the general study of TIDs and coherencies of $\overline{K}_{12}^2 > 0.50$ are found to be significant. If all three values of Δt fulfil the coherency criterion then the two time delays with greatest coherencies are employed to calculate an apparent propagation velocity.

The most important parameter to be determined directly from HF Doppler data is the time delay, Δt, between perturbations on the different radio paths. The magnitude of the time delays is not only employed to determine horizontal velocities, but also to differentiate between travelling and non-travelling disturbances. The time delays can be determined from either the cross correlation or cross spectral analysis. The cross correlation technique has several disadvantages which are not removed even when digital filtering is applied to the data sets and major difficulties occur for long period waves when data sets are short. The cross spectral analysis is a more powerful technique since it examines the properties of different frequency components without the need for filtering. Spectral estimates are improved by smoothing in the frequency domain, which suppresses leakage and reduces variance. The Hanning function can be conveniently adopted for this purpose. Rejection criteria for velocity estimates, based on coherency have been developed and are generally available in the literature.

Summary

The Doppler technique provides a simple and effective method for monitoring transient changes in the ionospheric electron density distribution. It has been extensively employed for studies of TIDs and by comparing the time delays produced by these disturbances on three or more spaced propagation paths, an estimate of their speed and direction can be obtained. A recent study using this technique as part of the Worldwide Atmospheric Gravity-wave Study (WAGS) has been published by Crowley and McCrea (1988). There are inherent difficulties in the measurement due mainly to the dispersive nature of the acoustic-gravity waves responsible for the TIDs, the presence of more than one wave in the observing region at any given time, and the problems involved in determining the time delays and the precise location of the reflection points. In spite of these difficulties, many of the important characteristics of TIDs have been established from Doppler observations. The low cost and the ability to monitor continuously over long time periods, renders the Doppler method an ideal tool for synoptic studies of thermospheric waves during the forthcoming ISTP period.

References

Blackman and Tukey, 1958, "The Measurement of Power Spectra". Dover, New York.

Crowley G and McCrea I W, 1988. Radio Science 23, 905.

Davies K, 1965, "Ionospheric Radio Propagation" 11 US National Bureau of Standards.

Davies K, 1968, AGARD Lecture Series No. 29.

Hines C O, 1974. J. Atmos.Terr.Phys. 36, 1179

Jenkins and Watts, 1968, "Spectral Analysis and its Applications." Holden-Day, London.

Jones D and Maude A D, 1968. J. Atmos.Terr.Phys. 30, 1487

Jones D and Maude A D, 1972, J. Atmos.Terr.Phys. 34, 1241

Sloan, 1969, IEEE. Trans.Audio and Electro acoust. AU-17, 133.

Tick, 1967, "Spectral Analysis of Time Series". Wiley, London.

Yuen C K, 1977. Proc. IEEE. 65, 984.

Chapter 15

THEORETICAL AND SIMULATION METHODS APPLIED TO HIGH LATITUDE, F REGION TURBULENCE

J.D. Huba

Space Plasma Branch
Plasma Physics Division
Naval Research Laboratory
Washington, DC 20375-5000

ABSTRACT

A general overview of the theoretical and simulation methods used to study plasma instabilities in the high latitude ionosphere is presented. This is done by way of example, using the $\mathbf{E} \times \mathbf{B}$ instability to illustrate the various techniques. The linear theory is presented from both the fluid and kinetic points of view, as well as both local theory and nonlocal theory. The nonlinear theories of the instability are also discussed, describing both analytical techniques and simulation methods (e.g., pseudo-spectral, finite difference). Finally, we compare the theoretical results with observational data which provides strong evidence that the $\mathbf{E} \times \mathbf{B}$ instability is active in the high latitude F region.

I. INTRODUCTION

The high latitude ionosphere is one of the more exciting regions of the near-earth space environment. This is primarily because it is "connected" to the distant magnetosphere and solar wind, and many phenomena which occur in the high latitude ionosphere are directly related to processes in the magnetosphere and solar wind, perhaps the most spectacular being auroral displays because of their visual impact. In recent years, high latitude plasma turbulence (i.e., electron density irregularities) has received considerable attention in the space plasma physics community [Keskinen and Ossakov, 1983; Kintner and Seyler, 1985; Tsunoda, 1988]. It is known that the scale sizes of plasma irregularities span an enormous range [Kelley et al., 1982; Baker et al., 1983; Rodriquez and Szuszczewicz, 1984], from a few centimeters to thousands of kilometers: eight orders of magnitude! It is clear that a variety of mechanisms are involved in creating this range of structure sizes. The types of mechanisms that have been proposed to explain irregularity production are electron precipitation, plasma mixing by the global convection pattern, plasma and fluid instabilities, etc. Thus, the subject is very interesting, in part, because of its complexity and the interplay of different physical processes. For a recent review of the subject, the interested reader is referred to a recent article by Tsunoda (1988).

There is reasonably good evidence that plasma irregularities with scale sizes $\lesssim 10$ km are produced primarily by an instability process. Of course, there still remains a large range of scale sizes to be explained. No single instability appears capable of generating all of the observed structure so a number of instabilities have been suggested: $\mathbf{E} \times \mathbf{B}$, Kelvin-Helmholtz, current convective, drift wave, Farley-Buneman, ion cyclotron, ion acoustic, lower hybrid, velocity shear, etc. Within the context of investigating a particular instability and its role in producing ionospheric turbulence, a number of questions should be answered at the outset to clarify the type of analysis to be used and its purpose. What is the equilibrium state of the system? Should fluid theory be used or kinetic theory? Should a local theory be developed or a nonlocal theory? Should a linear analysis be performed or a nonlinear analysis? What analytical and/or numerical methods are appropriate? The experimental data that one hopes to understand and explain provide the key to answering these questions. It is important to identify the ambient plasma and field conditions, the sources of free energy which can drive a plasma instability, the length and time scales associated with the fluctuations, etc.

The purpose of this paper is to provide an overview of the theoretical and computational methods used to study plasma instabilities in the high latitude ionosphere. We will focus on the $\mathbf{E} \times \mathbf{B}$ instability to illustrate the various techniques; this instability is also known as the gradient drift instability and the $\mathbf{E} \times \mathbf{B}$ gradient drift instability. The reasons for choosing this instability are as follows. First, the instability is relatively simple to understand and to model. Second, virtually all of the methods available to study a plasma instability have been applied to the $\mathbf{E} \times \mathbf{B}$ instability (to my knowledge, the only notable exception are particle and hybrid code simulations). And finally, there is good observational evidence that this instability is one of the causes of plasma structure in the high latitude ionosphere for scale sizes 100's m to 10's km. We add that this paper is very similar to a recent article by Huba (1989), but that we have expanded the analysis in this paper to include ion inertia effects and magnetospheric coupling.

The organization of the paper is as follows. In the next section we present a brief discussion of the ambient plasma and field conditions in the high latitude ionosphere, and provide a physical setting for the instability. In Section III we describe the linear theory of the instability. We present both the fluid theory and the kinetic theory, as well as both the local theory and the nonlocal theory. In Section IV we describe the nonlinear theory of the instability, emphasizing different analytical and numerical methods that have been developed. In Section V we present a discussion of the observational evidence for the $\mathbf{E} \times \mathbf{B}$ instability in the high latitude ionosphere, and compare the data with recent simulation results. In Section VI, we summarize the major points of the paper in the final section. We include an Appendix which describes an instability model which incorporates magnetospheric coupling.

II. PHYSICAL MODEL AND EQUILIBRIUM

It is important to first get some feeling for the physical parameters of the high latitude ionosphere, and how they relate to fundamental plasma parameters. In Table I we present typical values for various parameters. In general, one can characterize the high latitude F region as being a weakly collisional, low β, cold plasma.

In Fig. 1 we specify a physical situation which can occur in the high latitude F region and which can lead to the $\mathbf{E} \times \mathbf{B}$ instability [Rino et al., 1978; Vickrey et al., 1980]. We show F region "blob": this is a region of enhanced electron density (a factor of 2-10 above ambient) in the altitude regime 200-600 km with a longitudinal extent of ~ 100's km and a latitudinal extent of 1000's km. We also include an eastward ambient electric field $\mathbf{E}_0$ and note that there is a poleward density gradient on the equator side of the blob.

In Fig. 2 we simplify the physical model in Fig. 1 using a slab coordinate system. We take the ambient magnetic field to be in the z-direction ($\mathbf{B} = B_0\, e_z$), the ambient electric field to be in the y-direction ($\mathbf{E} = E_0\, e_y$), and the density to be inhomogeneous in the x-direction ($n_0 = n_0(x)$).

The basic equations describing the plasma are the following:

Table I: Typical F Region Parameters

Electron density:	$n_e \approx 10^3 - 10^6 \text{ cm}^{-3}$
Neutral density:	$n_n \lesssim 10^{12} \text{ cm}^{-3}$
Temperature:	$T_e \sim T_i \lesssim 0.1$ ev
Magnetic field:	$B_0 = 0.3$ G
Oxygen plasma:	$m_i = 16\ m_p$
Debye length:	$\lambda_{De} \approx 2.7$ mm
Electron gyroradius:	$\rho_e \approx 2.5$ cm
Ion gyroradius:	$\rho_i \approx 4.3$ m
Electron plasma frequency:	$\omega_{pe} \lesssim 5.6 \times 10^7 \text{ sec}^{-1}$
Electron cyclotron frequency:	$\Omega_e \approx 5.3 \times 10^6 \text{ sec}^{-1}$
Electron collision frequency:	$\nu_e \lesssim 1.4 \times 10^3 \text{ sec}^{-1}$
Ion plasma frequency:	$\omega_{pi} \lesssim 3.3 \times 10^5 \text{ sec}^{-1}$
Ion cyclotron frequency:	$\Omega_i \approx 1.8 \times 10^2 \text{ sec}^{-1}$
Ion collision frequency:	$\nu_i \lesssim 7 \text{ sec}^{-1}$
Plasma beta (β):	$\beta \lesssim 4.5 \times 10^{-5}$

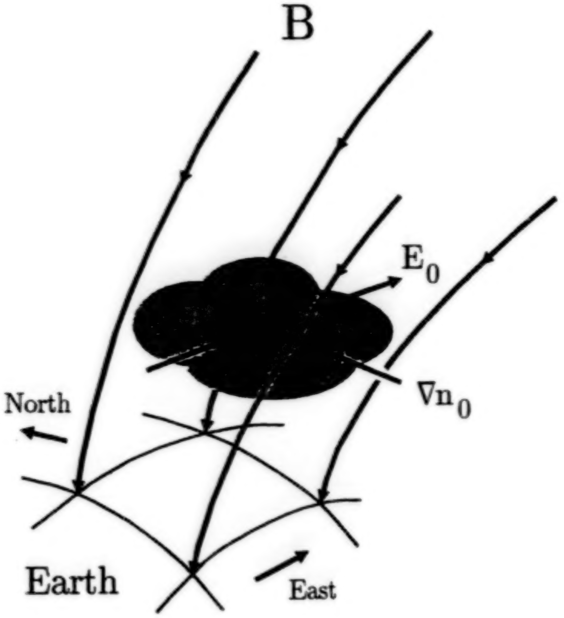


Fig. 1) Schematic showing a high latitude plasma blob. The ambient electric and magnetic fields, and density gradient are oriented so that the **E** x **B** instability can be excited.

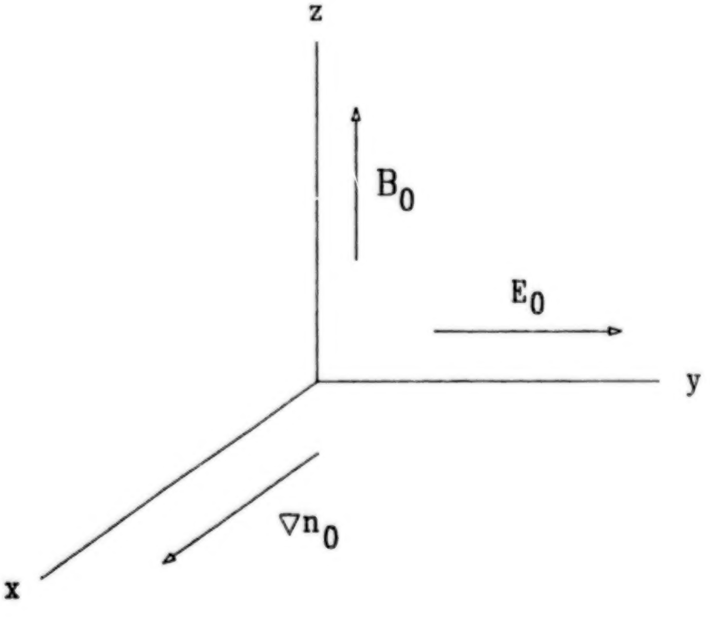


Fig. 2) Slab geometry and plasma configuration used in the analysis.

$$\frac{\partial n_\alpha}{\partial t} + \nabla\cdot n\underset{\sim}{V}_\alpha = 0 \tag{1}$$

$$\frac{d\underset{\sim}{V}_\alpha}{dt} = \frac{e_\alpha}{m_\alpha}\left(\underset{\sim}{E} + \frac{1}{c}\underset{\sim}{V}_\alpha \times \underset{\sim}{B}\right) - \nu_{\alpha n}V_\alpha \tag{2}$$

$$\nabla\cdot \underset{\sim}{J} = 0 \tag{3}$$

where α denotes species (i: ions, e: electrons) and the symbols have their usual meaning. In (2) $\nu_{\alpha n}$ is the particle-neutral collision frequency, we are in the neutral wind frame of reference (i.e., $\underset{\sim}{V}_n = 0$), and we have taken $T = 0$ for simplicity.

The equilibrium drifts are obtained from (2) by setting $d/dt = 0$ and solving for $V_{\alpha 0}$. We obtain

$$\underset{\sim}{V}_{\alpha 0} = \frac{cE_0}{B_0}\hat{e}_x + \frac{\nu_{\alpha n}}{\Omega_\alpha}\frac{cE_0}{B_0}\hat{e}_y \tag{4}$$

The first term in (4) is the usual E x B drift while the second term is the Pedersen drift. Note that the Pedersen drift, unlike the E x B drift, is specie dependent so that it gives rise to the Pedersen current. This current is the source of free energy which drives the $\underset{\sim}{E} \times \underset{\sim}{B}$ instability.

The equilibrium density is obtained from (1) by setting $\partial/\partial t = 0$. Thus, we require

$$\nabla\cdot n_{\alpha 0}(x)\ \underset{\sim}{V}_{\alpha 0} = 0 \tag{5}$$

which leads to $n_{\alpha 0}(x)V_{\alpha 0x}$ = constant. We choose to work in the $\underset{\sim}{E} \times \underset{\sim}{B}$ frame of reference, (i.e., $V_{\alpha 0x} = V_{\alpha 0x} - cE_0/B_0$) so that $V_{\alpha 0x} = 0$. Thus, the equilibrium density $n_{\alpha 0}(x)$ can be an arbitrary function of x.

III. LINEAR THEORY

A. Physical Picture

A simple picture of the physical processes involved in the $\underset{\sim}{E} \times \underset{\sim}{B}$ instability is illustrated in Fig. 3. We use the same geometry shown in Fig. 2 but are looking along the z-axis. In Fig. 3 we imagine a "heavy" fluid on top of a "light" fluid; the boundary between these two regions is denoted by the horizontal line. We perturb this boundary with a sinusoidal perturbation. In the F-region the ion Pedersen drift is much larger than the electron Pedersen drift because $\nu_{in}/\Omega_i \gg \nu_{en}/\Omega_e$. Thus, the ions (denoted by the solid line in Fig. 3) drift in the direction of $\underset{\sim}{E}_0$ and leave behind the electrons (denoted by the dashed line in Fig. 3). Because of the charge imbalance that results, a perturbed electrostatic field is set up, as shown. The perturbed electric field $\delta\underset{\sim}{E}$ then causes the plasma to drift with a velocity $\delta\underset{\sim}{V} = c\delta E/B\ e_x$. For the directions chosen for ∇n_0 and $\underset{\sim}{E}$, the perturbed velocity acts to enhance the density perturbation: the heavy

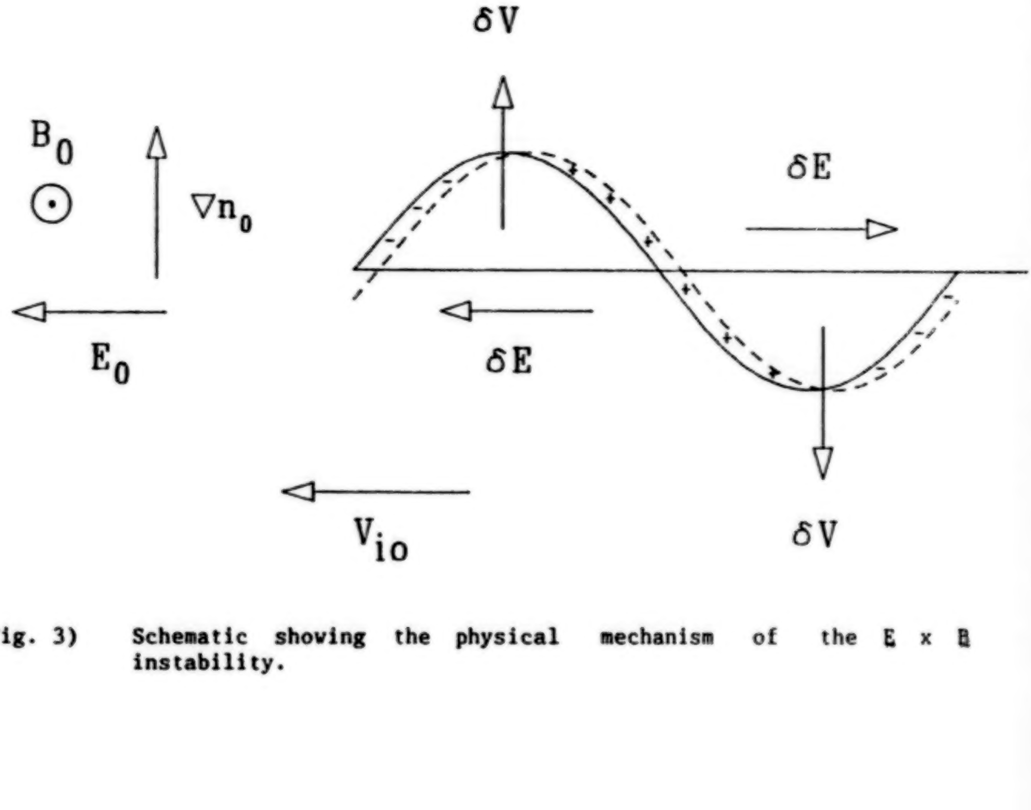


Fig. 3) Schematic showing the physical mechanism of the $\mathbf{E} \times \mathbf{B}$ instability.

fluid falls and the light fluid rises. This is characteristic of an interchange instability, such as the classic Rayleigh-Taylor instability. On the other hand, if the density gradient were chosen in the opposite direction, the perturbation would not grow. Thus, instability will only occur for $\mathbf{E}_0 \times \mathbf{e}_z \cdot \nabla n_0 > 0$.

B. Fluid Theory

Fluid theory is a very useful and important method to describe plasma dynamics, e.g., the onset and evolution of plasma instabilities. However, one must always keep in mind the limitations of fluid theory and not apply it to plasma regimes where it is not valid. In general, fluid theory is an appropriate theory to describe a plasma species α when kinetic effects are not important, e.g., in the absence of wave-particle interactions ($\omega \sim kv_\alpha$), finite Larmor radius effects ($kr_{L\alpha} \sim 1$) where v_α is the thermal velocity and $r_{L\alpha}$ is the gyroradius of species α, etc. Thus, one usually assumes the cold plasma, long wavelength approximation ($\omega \gg kv_\alpha$ and $kr_{L\alpha} \ll 1$) when using fluid theory.

With regard to the $\mathbf{E} \times \mathbf{B}$ instability in the ionosphere, fluid theory is valid for $kr_{Li} \ll 1$ and $\omega \ll \Omega_i$. This translates into the following spatial and temporal time scales: $\lambda \gg 25$ m and $\tau \gg 5 \times 10^{-3}$ sec. We note that the instability is a flute mode (i.e., $\mathbf{k} \cdot \mathbf{B}_0 = 0$); there are no wave-particle resonances transverse to B when $\omega \ll \Omega_i$ (more specifically, $\gamma < \Omega_i$ where γ is the growth rate).

We now derive a dispersion equation which describes the $\mathbf{E} \times \mathbf{B}$ instability in the fluid limit. For simplicity, we make the following assumptions: $kr_{Li} \ll 1$ (long wavelength), $\omega \ll \Omega_i$ (low frequency), $n_e = n_i$ (quasi-neutrality), $T = 0$ (cold plasma), $\mathbf{k} \cdot \mathbf{B} = 0$ (flute mode), $\beta \ll 1$ (low beta), and consider only electrostatic perturbations. In contrast to [Huba, 1989], we include the effects of ion inertia; we allow the ratio ν_{in}/ω to be arbitrary. The basic equations are the following:

$$\frac{\partial n}{\partial t} + \nabla \cdot n\mathbf{V}_e = 0 \tag{6}$$

$$0 = -\frac{e}{m_e}\left(\mathbf{E} + \frac{1}{c}\mathbf{V}_e \times \mathbf{B}_0\right) \tag{7}$$

$$\frac{d\mathbf{V}_i}{dt} = \frac{e}{m_i}\left(\mathbf{E} + \frac{1}{c}\mathbf{V}_i \times \mathbf{B}_0\right) - \nu_{in}\mathbf{V}_i \tag{8}$$

$$\nabla \cdot \mathbf{J} = \nabla \cdot \left[n\left(\mathbf{V}_i - \mathbf{V}_e\right)\right] = 0 \tag{9}$$

We perturb these equations about the equilibrium, i.e., we take $n = n_0(x) + \tilde{n}$, $\mathbf{E} = \mathbf{E}_0 - \nabla\tilde{\Phi}$, and $\mathbf{V}_\alpha = \mathbf{V}_{\alpha 0} + \tilde{\mathbf{V}}_\alpha$ where $\tilde{\Phi}$ is the perturbed electrostatic potential. The equilibrium drifts are $\mathbf{V}_{e0} = 0$ and $\mathbf{V}_{i0} = (\nu_{in}/\Omega_i)(cE_0/B_0)\mathbf{e}_y$. We take the perturbations to have the following form: $\tilde{p} = \tilde{p}(x)\exp[i(k_y y - \omega t)]$.

Substituting these quantities into (6) - (9) one obtains six equations with six unknowns, $(\tilde{n}, \tilde{\Phi}, \tilde{V}_{ex}, \tilde{V}_{ey}, \tilde{V}_{ix}, \tilde{V}_{iy})$. After some straightforward algebra these equations can be combined to yield a simple equation for $\tilde{\Phi}$:

$$\frac{\partial}{\partial x}\left[n_0 \frac{\partial\tilde{\phi}}{\partial x}\right] - n_0 k_y^2\left[1 + \frac{\nu_{in}}{\omega}\,\frac{1}{\omega + i\nu_{in}}\,\frac{cE_0}{B_0}\,\frac{n_0'}{n_0}\right]\tilde{\phi} = 0 \tag{10}$$

where $n_0' = \partial n_0/\partial x$. Interestingly, in the limit of strong collisions (i.e., $\nu_{in} \gg \omega$), we note that the mode equation is independent of the ion neutral collision frequency.

(1) Local Theory. In general, one usually solves (10) using the local approximation. This approximation, also referred to as the short wavelength limit, is based on the assumption that $\lambda \ll L_n$ where λ is the wavelength of the perturbation and L_n $(= (\partial \ln(n_0)/\partial x)^{-1})$ is the density gradient scale length. This is shown schematically in Fig. 4a; the plasma fluctuations are on a much smaller scale than the scale of associated with the change in ambient density.

The advantage of the local approximation is that one assumes that the perturbations are Fourier modes even in the inhomogeneous direction, i.e., $\tilde{p} \propto \exp[i(k_x x + k_y y - i\omega t)]$. If one further assumes that $k_y L_n \gg k_x L_n \gg 1$ then the first term in (10) can be neglected, and the dispersion equation is simply given by [Ossakov et al., 1978]

$$\omega(\omega + i\nu_{in}) + \nu_{in}\frac{V_0}{L_n} = 0 \tag{11}$$

where $V_0 = cE_0/B_0$ and $\omega = \omega_r + i\gamma$. We evaluate $1/L_n = n_0'/n_0$ at some $x = x_0$, typically where maximum growth occurs. Equation (11) can be simplified by considering the collisional and inertial limits, i.e., $\nu_{in} \gg \omega$ and $\nu_{in} \ll \omega$, respectively. It is easily shown that

$$\gamma = \begin{cases} V_0/L_n & ;\ \nu_{in} \gg \omega \ \text{(collisional)} \\ (V_0/L_n\nu_{in})^{1/2} & ;\ \nu_{in} \ll \omega \ \text{(inertial)} \end{cases} \tag{12}$$

From (12) we note that (1) instability only occurs when V_0 and L_n (or E_0 and n_0') are both positive or negative (i.e., in general, $\mathbf{E}_0 \times \mathbf{e}_z \cdot \nabla n_0 > 0$), (2) the mode is purely growing (i.e., $\omega_r = 0$), (3) the growth rate is independent of wavenumber k, and (4) the growth rate is largest in the collisional limit.

(2) Nonlocal Theory. The other analytical method for solving (10) is based on nonlocal theory. In this case one assumes $\lambda \gg L_n$, i.e., the long wavelength approximation [Huba and Zalesak, 1983; 1984]. This is shown schematically in Fig. 4b. Here, the plasma fluctuation scale length is much longer than the scale associated with the change in ambient density (e.g., surface waves (9)).

For this situation one must solve the differential equation (10). The easiest way is to assume a density profile given by a step function, i.e.,

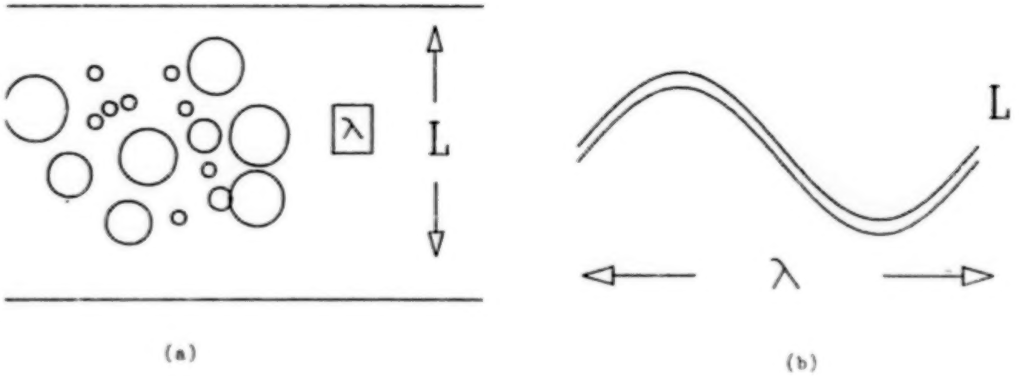


Fig. 4) Schematic showing the distinction between the (a) short wavelength limit and the (b) long wavelength limit.

$$n_0(x) = \begin{cases} n_2 & x > 0 \\ n_1 & x < 0 \end{cases} \tag{13}$$

The perturbed potential is assumed to be

$$\tilde{\phi}(x) = \begin{cases} \phi_2\, e^{-k_y x} & x > 0 \\ \phi_1\, e^{k_y x} & x < 0 \end{cases} \tag{14}$$

Integrating (10) across the discontinuous density layer at x = 0 yields the following matching conditions:

$$n_0 \frac{\partial \tilde{\phi}}{\partial x} \Big|_{-\varepsilon}^{\varepsilon} = - ik_y^2 \frac{cE_0}{B_0} \frac{1}{\omega} n_0 \Big|_{-\varepsilon}^{\varepsilon} \tag{15a}$$

$$\tilde{\phi}\Big|_{-\varepsilon}^{\varepsilon} = 0 \tag{15b}$$

where we let $\varepsilon \to 0$.

The dispersion equation is obtained by substituting (13) and (14) in the jump conditions given by (15). We then obtain the following dispersion equation

$$\omega(\omega + i\nu_{in}) + \nu_{in} k_y V_0 \frac{n_2 - n_1}{n_2 + n_1} = 0 \quad . \tag{16}$$

Note that (16) is similar to (11); one can recover (16) from (11) by making the substitution $1/L_n = k_y(n_2 - n_1)/(n_2 + n_1)$. The growth rate is easily obtained in the collisional and inertial limits, as in local theory case:

$$\gamma = \begin{cases} k_y V_0 \dfrac{n_2 - n_1}{n_2 + n_1} & ;\ \nu_{in} \gg \omega \quad \text{(collisional)} \\[2ex] \left[k_y V_0 \nu_{in} \dfrac{n_2 - n_1}{n_2 + n_1}\right]^{1/2} & ;\ \nu_{in} \ll \omega \quad \text{(inertial)} \end{cases} \tag{17}$$

As in local theory, instability only occurs when V_0 and $(n_2 - n_1)$ are both positive or negative, the mode is purely growing, and the collisional growth rate is larger than the inertial growth rate. However, unlike local theory, the growth rate in the long wavelength limit is proportional to k_y in the collisional limit and to $k_y^{1/2}$ in the inertial limit.

(3) General Picture. In the previous two sections we have presented relatively simple techniques to evaluate (10) analytically. These methods are very useful because they permit analytical solutions which are important in understanding the turn-on criteria associated with the instability, quantitative estimates of the growth rate, and how the unstable modes scale with different parameters (e.g., V_0, k_y, L_n, ...). In general, local theory is usually used (at least initially) because it is easy and because, in general, it yields the largest growth rate. On the other hand, it is important to also consider the nonlocal limit because large scale disruptions of plasma are often observed in which $\lambda > L_n$. Furthermore, some modes are only unstable in the nonlocal limit (e.g., the Kelvin-Helmholtz instability driven by sheared velocity flows transverse to B [Chandrasekhar, 1961]; applications to the high latitude ionosphere are given in Kintner and Seyler (1985) and Keskinen et al. (1988)).

Ultimately, if one wishes to determine the growth rate in the regime $\lambda \sim L_n$ one must solve (10) numerically for a specific density profile. We sketch out a typical solution in Fig. 5 for the $\underset{\sim}{E} \times \underset{\sim}{B}$ instability in the collisional limit. The solid line is the analytical solution given by (11) and (15), and the dashed line is the numerical solution. As expected, the numerical solution asymptotes to the appropriate limits. However, for $k_y L_n$ sufficiently large, i.e., very small wavelengths, additional physical effects enter the picture, such as diffusion and finite Larmor radius effects which have not been included in the analysis.

For the sake of completeness, we write down the appropriate momentum transfer equations which include these additional effects. We do not solve them here but simply give references where they have been solved. The electron momentum transfer equation can be written as

$$0 = -\frac{e}{m_e}\left(\underset{\sim}{E} \quad \frac{1}{c}\underset{\sim}{V}_e \times \underset{\sim}{B}\right) - \nu_{en}\underset{\sim}{V}_e - \nu_{ei}\left(\underset{\sim}{V}_e - \underset{\sim}{V}_i\right) - \frac{\nabla P_e}{n m_e} \tag{18}$$

where $P_e = nT_e$. The second and third terms are the collisional drags with neutrals and ions, respectively, and the final term is electron pressure. The combination of collisional drag and pressure results in diffusion damping of the $\underset{\sim}{E} \times \underset{\sim}{B}$ instability.

The ion momentum transfer equation can be written as

$$\frac{d\underset{\sim}{V}_i}{dt} = \frac{e}{m_i}\left(\underset{\sim}{E} + \frac{1}{c}\underset{\sim}{V}_i \times \underset{\sim}{B}\right) - \nu_{in}\underset{\sim}{V}_i - \nu_{ie}\left(\underset{\sim}{V}_i - \underset{\sim}{V}_e\right) - \frac{\nabla P_i}{n m_i} - \nabla \cdot \underset{\approx}{\Pi}_i \tag{19}$$

which is somewhat more complicated than the electron equation. The additional terms here are (1) collisional drag with the electrons, (2) ion pressure and (3) the ion stress tensor $\underset{\approx}{\Pi}_i$. The ion stress tensor includes the effects of ion-ion collisions and finite Larmor radius corrections. In general, all of the additional terms in (18) and (19) lead to damping (or reduced growth) of the $\underset{\sim}{E} \times \underset{\sim}{B}$ instability [Sperling and Glassman, 1985].

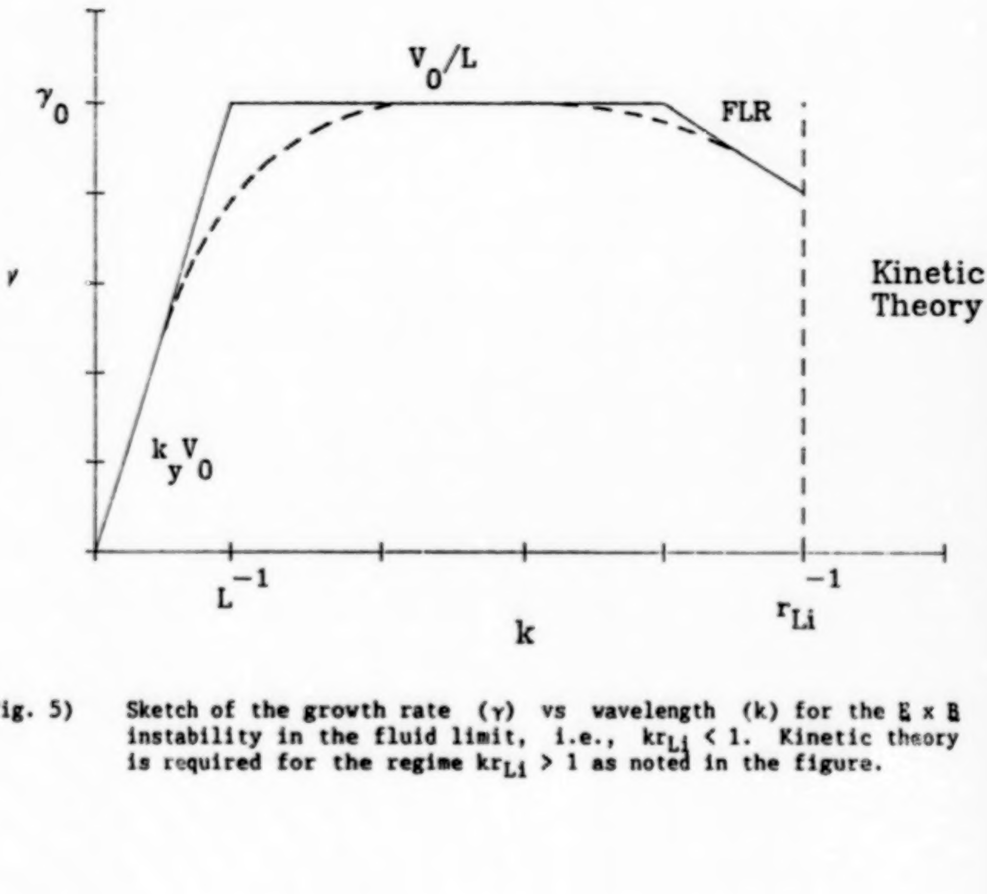


Fig. 5) Sketch of the growth rate (γ) vs wavelength (k) for the E x B instability in the fluid limit, i.e., $kr_{Li} < 1$. Kinetic theory is required for the regime $kr_{Li} > 1$ as noted in the figure.

C. Kinetic Theory

As indicated in the previous section, fluid theory has limitations. For the case of the $\underset{\sim}{E} \times \underset{\sim}{B}$ instability, the theory is restricted to the regime $kr_{Li} \ll 1$. When $kr_{Li} \sim O(1)$ the finite Larmor radius effects become important which cannot be handled by fluid theory. One can use the appropriate ion stress tensor in the fluid limit [Roberts and Taylor, 1962] to gain some understanding of how finite Larmor radius effects influence the instability but one still requires $k_y r_{Li}$ be a small parameter. This is shown schematically in Fig. 5 (labelled FLR). The best (and correct) way to handle the problem is to use kinetic theory: solve the Vlasov equation with an appropriate collision model. This method removes all wavelength and temperature restrictions.

To my knowledge, very little work has been done on the kinetic theory of the $\underset{\sim}{E} \times \underset{\sim}{B}$ instability. This is because observations of plasma structure have focused on the wavelength regimes 100's m - 10's km so that fluid theory is more than adequate. Gary and Cole (1983) have presented a kinetic theory of the $\underset{\sim}{E} \times \underset{\sim}{B}$ instability; however, their formulation is questionable because they did not consider the ion Pedersen drift in the orbit equations. We now outline a kinetic theory of the $\underset{\sim}{E} \times \underset{\sim}{B}$ instability; a detailed analysis can be found in Huba et al. (1989).

The starting point of kinetic theory is the Vlasov equation using the Krook collision model:

$$\frac{\partial f_\alpha}{\partial t} + \underset{\sim}{v} \cdot \frac{\partial f_\alpha}{\partial \underset{\sim}{x}} + \frac{e_\alpha}{m_\alpha} \left[\underset{\sim}{E} + \frac{1}{c} \underset{\sim}{v} \times \underset{\sim}{B} \right] \cdot \frac{\partial f_\alpha}{\partial \underset{\sim}{v}} = - \nu_{\alpha n} \left[f_\alpha - \frac{\tilde{n}_\alpha}{n_\alpha} f_\alpha^M \right] \tag{20}$$

where f_α is the distribution function, f_α^M is a Maxwellian distribution function, $\nu_{\alpha n}$ is the collision frequency with neutrals, and $\tilde{n}_\alpha$ is the perturbed density (all of species α).

The first step in solving (20) is to determine the zeroth-order distribution function $f_\alpha^0(x, \underset{\sim}{v})$ (i.e., the equilibrium distribution function). The equilibrium distribution function for the electrons and ions can be written as functions of the single-particle constants of motion in the unperturbed fields. For electrons we take

$$F_{e0}(v_\perp{}^2, v_z{}^2, X) = n_0(X)(\pi v_e{}^2)^{-3/2} \exp[(v_\perp{}^2 + v_z{}^2)/v_e{}^2] \tag{21}$$

where $v_\perp{}^2 = v_x{}^2 + (v_y - V_0)^2$, $X = x - (v_y - V_0)/\Omega_e$, $V_0 = cE_0/B_0$, and $v_e = (2T_e/m_e)^{1/2}$. The unperturbed electron distribution function can be expanded locally about some $x = x_0$ based upon the weak inhomogeneity approximation which yields

$$F_{e0}(v_\perp{}^2, v_z{}^2) = n_0(\pi v_e{}^2)^{-3/2} \exp[-(v_\perp{}^2 + v_z{}^2)/v_e{}^2](1 - \frac{v_y - V_E}{\Omega_e L_n}) \tag{22}$$

where n_0 and L_n are evaluated locally. Note that we are using the local approximation (i.e., $\lambda \ll L_n$).

We assume a local, drifting Maxwellian distribution for the ions given by

$$F_{i0}(v_\perp{}^2, v_z{}^2) = n(\pi v_i{}^2)^{-3/2}\exp[-(v_\perp{}^2 + v_z{}^2)/v_i{}^2] \tag{23}$$

where $v_\perp{}^2 = v_x{}^2 + (v_y - V_{0y})^2$, $v_i = (2T_i/m_i)^{1/2}$, $V_{0y} = V_{di} + V_{pi}$, $V_{di} = v_i^2/\Omega_i L_n$ is the ion diamagnetic drift, and $V_{pi} = (\nu_{in}/\Omega_i)V_0$ is the ion Pedersen drift. The ion diamagnetic drift is proportional to the temperature T_i and was not included in the fluid analysis because we assumed $T_i = 0$.

The next step is to determine the perturbed distribution function f_α. This is usually done by integrating over the unperturbed orbits $x'(\tau)$ (i.e., the method of characteristics, see Krall and Trivelpiece (1973)). For the problem at hand we have

$$\begin{aligned}\delta f_\alpha = {} & -\frac{e_\alpha}{m_\alpha}\int_{-\infty}^{0} d\tau \exp[i\underset{\sim}{k} \cdot \underset{\sim}{x}'(\tau) - i(\omega + i\nu_{\alpha n})\tau]\, \nabla\delta\phi \cdot \frac{\partial F_{\alpha 0}}{\partial \underset{\sim}{v}'} \\ & + \nu_{\alpha n}\frac{\delta n_\alpha}{n}\int_{-\infty}^{0} d\tau \exp[ik \cdot x'(\tau) - i(\omega + i\nu_{\alpha n})\tau]F_{\alpha 0} \quad .\end{aligned} \tag{24}$$

Finally, after the perturbed distribution function is found, one then calculates the perturbed density fluctuation $\tilde{n}_\alpha = \int d^3v\, f_\alpha$ and substitutes it into Poisson's equation $\nabla^2\Phi = 4\pi(\tilde{n}_i - \tilde{n}_e)$ to determine the electrostatic dispersion equation. One finds that the dispersion equation is given by

$$D(\omega,k) = 1 + \chi_e + \chi_i = 0 \tag{25}$$

where

$$\chi_i = \frac{\omega_{pi}^2}{k^2 v_i^2}\left[1 - \frac{\left(\omega - k_y V_{di} - k_y V_{pi} + i\nu_{in}\right)G_i}{1 - i\nu_{in}G_i}\right]$$

$$\chi_e = \frac{\omega_{pe}^2}{k^2 v_e^2}\left[1 - \frac{\left(\omega - k_y V_{de} + i\nu_{in}\right)G_e}{1 - i\nu_{en}G_e}\right]$$

and $G_i = \exp(-b_i)I_0(b_i)/(\omega + k_y V_{pi} + i\nu_{in})$, $G_e = \exp(-b_e)I_0(b_e)/(\omega + i\nu_{en})$, $\omega_{p\alpha}^2 = 4\pi ne^2/m_\alpha$, $V_{d\alpha} = v_\alpha^2/\Omega_\alpha L_n$, $V_{pi} = (\nu_{in}/\Omega_\alpha)V_0$, $V_0 = cE_0/B_0$, $b_\alpha = k_y^2 v_\alpha^2/\Omega_\alpha^2$, and I_0 is the modified Bessel function of order 0. In writing χ_α we have assumed flute perturbation ($k \cdot B_0 = 0$), and have taken $\omega/\Omega_\alpha \ll 1$ and $\nu_{\alpha n}/\Omega_\alpha \ll 1$. In the limit $T_e = T_i = 0$ and $\nu_{en} = 0$, one can easily recover (11) from (25).

In Fig. 6 we show the growth rate as a function of kr_{Li} for the following parameters: $V_{di} = 0.01\, v_i$, $V_0 = 0.10\, v_i$, $\nu_{in} = 0.10\, \Omega_i$, $T_i = T_e$, $m_i = 1836\, m_e$, and $\nu_{en} = (0.0,\ 0.1,\ 1.0)\, \Omega_i$. Several interesting features are the following. First, as $kr_{Li} \rightarrow 1$ we note that the growth rate decreases in magnitude from its value when $kr_{Li} << 1$ (although it may seem that there is growth when $kr_{Li} = 0$ from Fig. 6, this is not the case as discussed in Section III.B.2). This is a manifestation of finite Larmor radius effects which have a stabilizing influence on the instability. Second, for $\nu_{en} < \nu_{in}$ it is found the growth rate actually maximizes in the regime $kr_{Li} >> 1$. This behavior is analogous to the kinetic interchange mode discussed in Gary and Thomsen (1982). And finally, for $\nu_{en} > \nu_{in}$ one finds that the mode is stable for sufficiently large kr_{Li}. This is caused by electron diffusion damping of the mode, i.e., $\gamma \simeq \gamma_0 - k^2D$ where $D = \nu_{en}r_{Le}^2$.

IV. NONLINEAR THEORY

Although linear theory is important in determining whether or not an instability can be excited in a specific physical situation (i.e., by providing turn-on criteria and estimates of growth rates), it is only the first step in developing a full understanding of plasma turbulence. The second step, and the more difficult one, involves the development of a nonlinear theory. This is crucial because experimental data are obtained, in general, during the nonlinear (saturated) phase of the instability. Thus, in order to meaningfully compare observational data with theoretical models, a nonlinear theory is clearly needed. For example, predictions of the saturated amplitudes of fluctuating quantities (e.g., $|\tilde{n}|^2$, $|E|^2$) and power spectral densities (e.g., $|\tilde{n}/n_0|^2$ vs k) are useful in making comparisons to the types of data obtained in high latitude experimental campaigns.

As noted above, developing a nonlinear theory is usually much harder than developing a linear theory. Analytical studies are possible and helpful, but usually require a number of simplifying approximations and assumptions in order to make any progress mathematically; unfortunately, these assumptions often break down when applying the theory to realistic ionospheric parameters. To overcome this problem, one can resort to nonlinear numerical simulations which can remove many of the limitations of an analytical theory. However, one must then contend with the subtleties of numerical analysis which is fraught with its own set of difficulties. Nevertheless, large scale computational simulation codes offer the best method to unravel the details of the nonlinear evolution of plasma instabilities.

In this section we will initially discuss analytical techniques used to study the nonlinear evolution of the $\mathbf{E} \times \mathbf{B}$ instability. The emphasis will be on the underlying physics of mode coupling. Following this we will present a discussion of nonlinear simulation studies of the instability.

A. Analytical Results

The major assumptions that are made in developing analytical models of the nonlinear behavior of the $\mathbf{E} \times \mathbf{B}$ instability are the following. First, it is assumed that the fluctuating quantities remain small, i.e., $\tilde{n} << n_0$ and $E << E_0$. Based on this assumption only quadratic nonlinearities are retained. Second, the local approximation is assumed; the perturbation wavelengths are small compared to equilibrium gradient scale lengths (e.g., $\lambda << L_n$ as shown in Fig. 4a). And finally, it is assumed that the density and field fluctuations do not modify the equilibrium. Thus, only

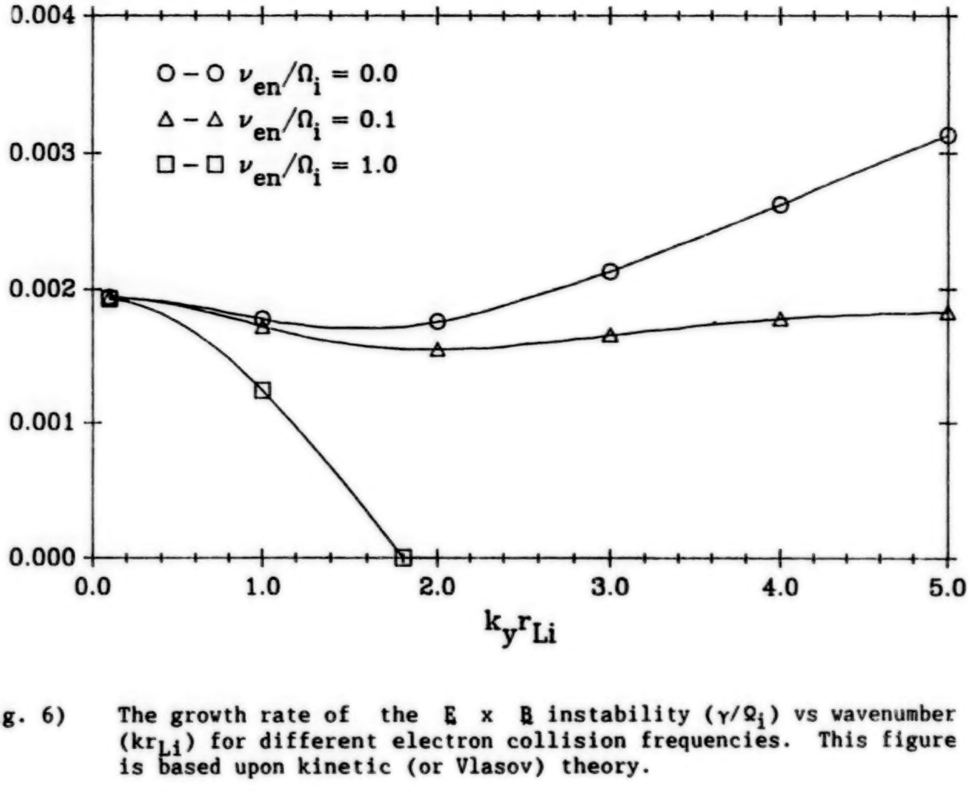


Fig. 6) The growth rate of the **E** x **B** instability (γ/Ω_i) vs wavenumber (kr_{Li}) for different electron collision frequencies. This figure is based upon kinetic (or Vlasov) theory.

perturbations need to be followed in time.

The underlying physical mechanism which saturates the $\mathbf{E} \times \mathbf{B}$ instability in the analyses to be discussed is mode coupling: linearly growing modes couple to linearly damped modes. A steady state is achieved when a balance is reached between the energy being put into growing modes and the energy being removed by some dissipation process. It is crucial that there be an energy sink in this model. Otherwise the instability would grow without bounds because the driver is being held constant. Thus, the equations presented in Section III.B (i.e., (6) - (9)) have to be modified in order to introduce damped modes. This is accomplished by replacing the electron momentum equation (7) with the following equation

$$0 = -\frac{e}{m_e}(\underset{\sim}{E} + \underset{\sim}{V}_e \times \underset{\sim}{B}/c) - \nu_{en}\underset{\sim}{V}_e - \frac{T_e}{m_e}\frac{\nabla n}{n} . \tag{26}$$

As indicated in Section III.B.3, the inclusion of electron-neutral collisional drag and finite electron temperature lead to diffusion damping of the instability. Following the analysis described in Section III, one can easily show that the dispersion equation can be written as

$$(\gamma + \nu_{in})(\gamma + k^2 D) = \frac{k_y^2}{k^2}\frac{\nu_{in}V_0}{L_n} \tag{27}$$

where $k^2 = k_x^2 + k_y^2$ and $D = \nu_{en}r_{Le}^2$ where $r_{Le} = v_e/\Omega_e$ and $v_e = (T_e/m_e)^{1/2}$. The first term in (24) drives the instability; it is similar to (11) but is modified to account for modes with finite k_x. The second term is the damping term; the density fluctuations are "smoothed out" by collisional diffusion. For example, in the collisional limit ($\nu_{in} >> \gamma$), it is easily shown that

$$\gamma = \frac{k_y^2}{k^2}\frac{V_0}{L_n} - k^2 D \tag{28}$$

Given this general background, we now present two examples of theoretical analyses which provide estimates of the saturated amplitudes of the density and field fluctuations, and of the nature of the power spectral density.

(1) Three Mode Coupling. A relatively simple analysis based upon a three mode interaction has been applied to a host of interchange instabilities in the E and F regions (e.g., gradient drift [Rognlien and Weinstock, 1974], collisional Rayleigh-Taylor [Chaturvedi and Ossakow, 1977], $\mathbf{E} \times \mathbf{B}$ [Chaturvedi and Ossakow, 1979], current convective [Chaturvedi and Ossakow, 1981], inertial interchange modes [Huba et al., 1985; Hassam et al., 1986]). This technique provides a simple understanding of the mode coupling process, and an estimate of the saturated fluctuation levels.

The equations used in the analysis are (6), (8), (9), and (26). Perturbing these equations and retaining only quadratic nonlinearities, we arrive at the following set of coupled nonlinear equations for the $\mathbf{E} \times \mathbf{B}$ instability,

$$\frac{\partial}{\partial t}\frac{\tilde{n}}{n_0} - \frac{c}{B}\nabla\tilde{\phi} \times \hat{e}_z \cdot \frac{\nabla n_0}{n_0} - D\nabla^2\frac{\tilde{n}}{n_0} = \frac{c}{Bn_0}\nabla\tilde{\phi} \times \hat{e}_z \cdot \nabla\tilde{n} \tag{29}$$

and

$$\left[\frac{\partial}{\partial t} + \hat{e}_z \times \nabla\tilde{\phi} \cdot \nabla + \nu_{in}\right]\nabla^2\phi = \nu_{in}E_0\,\hat{e}_y \cdot \nabla\frac{\tilde{n}}{n_0} \tag{30}$$

Equation (29) is the electron continuity equation. The second term on the LHS is the linear $\mathbf{E} \times \mathbf{B}$ drift, the third term is electron diffusion, and on the RHS is the nonlinear $\mathbf{E} \times \mathbf{B}$ drift term which is the only nonlinearity retained. Equation (30) comes from current conservation ($\nabla \cdot \mathbf{J} = 0$) and contains the driver (i.e., the Pedersen drift $\propto V_0$). Linearizing (29) and (30) yields the dispersion equation given by (27).

The nonlinear mode coupling can be seen by substituting the following density and potential fluctuations into (29) and (30): $\tilde{n} \propto \sin(x)\cos(y)$ and $\tilde{\Phi} \propto \sin(x)\sin(y)$. It is easily found that the $\mathbf{E} \times \mathbf{B}$ nonlinearity in (29) drives a density perturbation $\tilde{n} \propto \sin(2x)$. Thus, we will consider the following perturbations,

$$\tilde{\phi} = \tilde{\Phi}_{1,1}\sin(k_x x)\sin(k_y y) \tag{31}$$

$$\tilde{n} = \tilde{n}_{1,1}\sin(k_x x)\cos(k_y y) + \tilde{n}_{2,0}\sin(2k_x x) \tag{32}$$

Substituting (31) and (32) into (29) and (30) yields the following set of mode coupled equations,

$$\frac{\partial\tilde{\Phi}_{1,1}}{\partial t} = -\nu_{in}\tilde{\Phi}_{1,1} + \nu_{in}E_0\frac{k_y}{k^2}\frac{\tilde{n}_{1,1}}{n_0} \tag{33}$$

$$\frac{\partial\tilde{n}_{1,1}}{\partial t} = -k_y^2 D\,\tilde{n}_{1,1} + \frac{cn_0}{B}\frac{k_y}{L_n}\tilde{\Phi}_{1,1} - \frac{c}{Bn_0}k_x k_y\tilde{\Phi}_{1,1}\tilde{n}_{2,0} \tag{34}$$

$$\frac{\partial\tilde{n}_{2,0}}{\partial t} = -4k_x^2 D\,\tilde{n}_{2,0} + \frac{c}{B}\frac{1}{2n_0}k_x k_y\tilde{\Phi}_{1,1}\tilde{n}_{1,1} \tag{35}$$

The final terms on the RHS of (34) and (35) are the nonlinear coupling terms; note that the coupling leads to damping of the (1,1) mode but to growth of the (2,0) mode. Thus, when the (1,1) mode is linearly unstable (as determined by (27)), it eventually drives the (2,0) mode unstable when the (1,1) mode reaches a sufficiently large amplitude. This in turn causes the growth of the (1,1) mode to slow and eventually stop; ultimately a steady state is achieved. We show a schematic of the mode coupling in Fig.

7. Here, the boundary between growing and damped modes is given by the marginal stability criterion $\gamma = 0$.

We can estimate the saturation levels of the fluctuations by finding the steady state solution to (33) - (35). We let $\partial/\partial t = 0$ and solve for $\Phi_{1,1}$, $\tilde{n}_{1,1}$ and $\tilde{n}_{2,0}$. We find that for a weakly damped system that

$$\tilde{n}_{1,1} = \frac{2k}{k_y} \left(\frac{2D}{V_0 L_n}\right)^{1/2} n_0 \tag{36}$$

$$\tilde{n}_{2,0} = \frac{1}{k_x L_n} n_0 \tag{37}$$

$$\tilde{E}_{1,1} = 2\left(\frac{2D}{V_0 L_n}\right)^{1/2} E_0 \tag{38}$$

where we have taken $E_{1,1} \simeq -k\Phi_{1,1}$ and have assumed $V_0/L_n >> k^2 D$. We can quantify $\tilde{n}$ and E at saturation by using the following high latitude parameters: $V_0 = cE_0/B_0 = 100$ m/sec, $L_n = 10$ km, $k_x = k_y = 0.01$ m^{-1} ($\lambda_x = \lambda_y = 600$ m), $T_e = T_i = 0.1$ ev, $\nu_e = 1000$ sec^{-1}, $r_{Le} = 2.5$ cm, and $D = 1$ m^2/sec. We find that $\tilde{n}_{1,1}/n_0 \simeq 0.4\%$, $\tilde{n}_{2,0}/n_0 \simeq 1.0\%$, and $E_{1,1}/E_0 \simeq 0.3\%$. Thus, the instability can saturate at relatively low levels consistent with the assumption that the fluctuation level be small.

An interesting extension of this analysis has been described in Huba et al. (1985) in which it was demonstrated that the mode coupling equations (33) - (35) reduce to the famous Lorenz equations [Lorenz, 1963]. Thus, this three mode system can exhibit a strange attractor with chaotic behavior. Ion inertia plays a crucial role in this phenomenon in that if it is neglected then the three mode system does not exhibit chaos and a stable convection pattern results.

(2) Power Spectrum. Although the three mode calculation highlights the importance of nonlinear mode coupling and provides estimates of the saturated fluctuation levels, it is a gross simplification of the actual interchange process in the ionosphere because ionospheric turbulence is a many mode phenomenon. Clearly, numerical simulations are needed to follow the complex interaction of large numbers of modes. This will be discussed in the next section. However, an analytic estimate of the spatial power spectrum of the $\mathbf{E} \times \mathbf{B}$ instability can be obtained from conservation laws associated with the fundamental plasma fluid equations. As in the three mode system, the analysis is based upon the assumption that a steady state can be achieved via mode coupling; the growing modes are saturated by transferring energy to damped modes. The calculation that follows is from Keskinen and Ossakow (1981) and Keskinen (1989). The interested reader is referred to these articles for more details.

Equations (29) and (30) are the basic equations used in the analysis. We first multiply (29) by $\tilde{n}$ and integrate over all x and y. This leads to the following equation

$$\int dxdy \left[\frac{1}{2}\frac{\partial}{\partial t}|\tilde{n}|^2 - \tilde{n}\nabla\tilde{\phi} \times \hat{e}_z \cdot \nabla n_0 - \tilde{n}\nabla\tilde{\phi} \times \hat{e}_z \cdot \nabla\tilde{n} - D\tilde{n}\nabla^2\tilde{n}\right] = 0 \tag{39}$$

The third term in (39) vanishes upon integration over all x and y because $\tilde{n}$

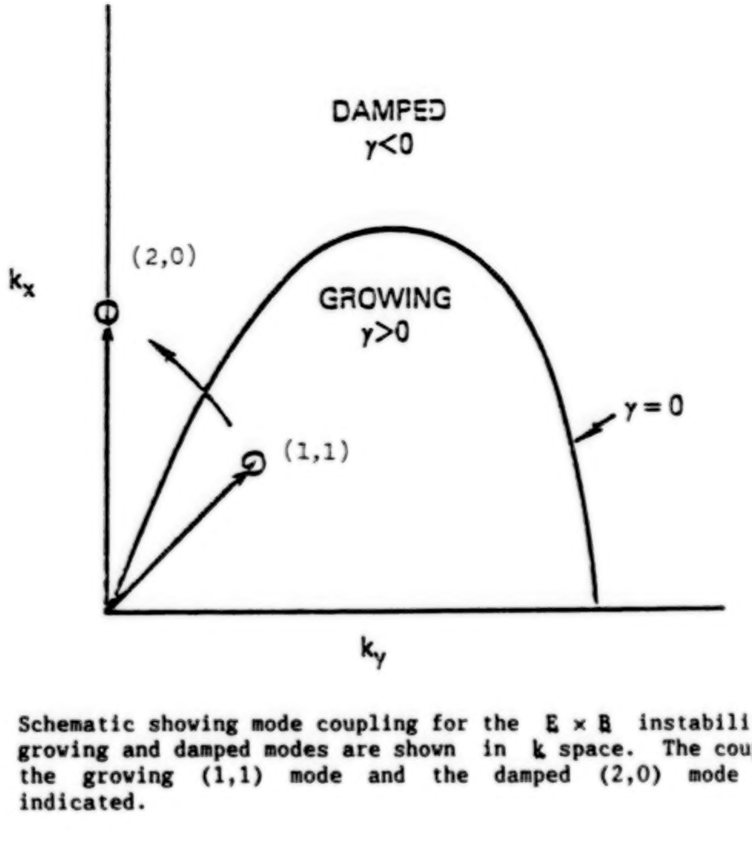


Fig. 7) Schematic showing mode coupling for the $\mathbf{E} \times \mathbf{B}$ instability. The growing and damped modes are shown in $\mathbf{k}$ space. The coupling of the growing (1,1) mode and the damped (2,0) mode is also indicated.

$\rightarrow 0$ as $x,y \rightarrow \infty$; this can be shown using Gauss' Law. We Fourier expand $\tilde{n}$ and $\tilde{\Phi}$ and obtain the following expression from (39)

$$\int dk_x dk_y \left[\frac{1}{2}\frac{\partial}{\partial t} I_{\underset{\sim}{k}} + i\frac{c}{B}\hat{e}_z \times \underset{\sim}{k} \cdot \nabla n_0\, \tilde{n}_{-\underset{\sim}{k}} \tilde{\phi}_{\underset{\sim}{k}} + Dk^2 I_{\underset{\sim}{k}}\right] = 0 \qquad (40)$$

where $I_{\underset{\sim}{k}} = |\tilde{n}_{\underset{\sim}{k}}|^2$. From (30) we obtain

$$\tilde{\phi}_{\underset{\sim}{k}} = \frac{ik_y E_0}{n_0 k^2}\frac{i\nu_{in}}{\omega + i\nu_{in}}\tilde{n}_{\underset{\sim}{k}} \qquad (41)$$

Substituting (41) into (40) we find that

$$\int dk_x dk_y \left[\frac{1}{2}\frac{\partial}{\partial t} I_{\underset{\sim}{k}} - \gamma_{\underset{\sim}{k}} I_{\underset{\sim}{k}}\right] = 0 \qquad (42)$$

where

$$\gamma_{\underset{\sim}{k}} = \frac{k_y^2}{k^2}\frac{V_0}{L_n} - k^2 D \qquad ; \ \nu_{in} \gg \gamma \ \text{(collisional)} \qquad (43a)$$

$$\gamma_{\underset{\sim}{k}} = \frac{k_y}{k}\left[\frac{\nu_{in} V_0}{L_n}\right]^{1/2} - \frac{k^2 D}{2} \qquad ; \ \nu_{in} \ll \gamma \ \text{(inertial)} \qquad (43b)$$

and we have used (27). Assuming a steady state and transforming to a cylindrical coordinate system, we find from (42) and (43) that

$$\int_{k_{min}}^{k_c} dk\, k \int_0^{2\pi} d\theta\, \gamma_{\underset{\sim}{k}}^{g} I_{\underset{\sim}{k}} = \int_{k_c}^{k_{max}} dk\, k \int_0^{2\pi} d\theta\, \gamma_{\underset{\sim}{k}}^{d} I_{\underset{\sim}{k}} \qquad (44)$$

where $\gamma_k{}^g$ is the positive term in (43) (i.e., the growth term), and $\gamma_k{}^d$ is the negative term in (43) (i.e., the damping term). In (43) it is assumed that the waves are linearly unstable in the range $k_{min} - k_c$, and are linearly damped in the range $k_c - k_{max}$; k_c is the wavenumber which corresponds to marginal stability ($\gamma = 0$).

It is now assumed that the spectrum $I(k,\theta)$ has the following form

$$I(k,\theta) = I_0 \cos^m\theta \, (1 + k^2/k_0^2)^{-(n+1)/2} \qquad (45)$$

where $m > 0$ and I_0 = constant. Substituting (45) into (44) it can be shown that

$$\left(\frac{k_0}{k_{max}}\right)^{n-1} = \frac{3-n}{n-1} \frac{\Gamma(\frac{m+2}{2})\Gamma(\frac{m+2}{2})}{\Gamma(\frac{m+1}{2})\Gamma(\frac{m+3}{2})} \frac{2(\nu_{in}V_0/L_n)^{1/2}}{k_{max}^2 D} \quad ; \quad \nu_{in} << \gamma \qquad (46a)$$

$$\left(\frac{k_0}{k_{max}}\right)^{n-1} = \frac{3-n}{n-1} \frac{\Gamma(\frac{m+1}{2})\Gamma(\frac{m+4}{2})}{\Gamma(\frac{m+2}{2})\Gamma(\frac{m+3}{2})} \frac{V_0/L_n}{k_{max}^2 D} \quad ; \quad \nu_{in} >> \gamma \qquad (46b)$$

where Γ is the gamma function and is defined as

$$\Gamma(z) = \int_0^\infty dt \; t^{z-1} e^{-t}$$

and it is assumed that $n > 1$ and $k_{min} < k_0 < k_c$. From (46) it is clear that $1 < n < 3$.

For an inverse power law with a spectral index $n = 2$ and $m = 2$ it is found that $k_0 \propto L_n^{-1}$. Equation (46a) is applicable for low altitudes (< 500 km) in the high latitude ionosphere. If we take $V_0 = 100$ m/sec, $D = 1$ m^2/sec, $L_n = 10$ km, and $k_{max} = 0.63$ m^{-1} ($\lambda_{min} = 10$ m), we find that $2\pi/k_0 = 500$ m which is consistent with observations [Tsunoda, 1989].

B. Numerical Results

As indicated earlier, large scale numerical simulations offer the best technique to understand the nonlinear evolution of plasma instabilities. In general, two types of simulations have been performed to study the nonlinear evolution of interchange instabilities. One type of simulation focuses on the evolution of small-scale turbulence, i.e., $kL_n >> 1$ (see Fig. 4a). The purpose of these simulations is to obtain an understanding of the cascade of wave energy from large to small wavelengths in two dimensions, and to calculate the spectral characteristics of the turbulence. Basically, these studies are an improvement of the analytic studies discussed above. The other type models the macroscopic evolution of the density layer (length scales $0.1 \lesssim kL_n \lesssim 3$). These simulations have generally been performed in configuration space using finite-difference techniques [Zalesak and Ossakow, 1980; Zalesak et al., 1985; Mitchell et al., 1985] although there have recent simulations using pseudo-spectral techniques [Zargham and Seyler, 1987; Kelley et al., 1987]. We now discuss in detail the important results from these types of simulations.

(1) Small-Scale Structure. We first discuss simulation studies of small-scale turbulence, i.e., $kL_n >> 1$. We study the steady state turbulence that develops in a model system described as follows. A weakly collisional, magnetized plasma ($\mathbf{B} = B_0 \mathbf{e}_z$) is confined between two horizontal conducting boundaries a distance L apart; the plasma density at the upper boundary ($x = L$) is higher than that at the lower boundary ($x = 0$). An ambient electric is in the y-direction, $\mathbf{E} = E_0 \mathbf{e}_y$. Thus, the system

is unstable to the $\mathbf{E} \times \mathbf{B}$ instability. This model is clearly not a complete representation of the ionospheric plasma; its chief advantages are that it is a well-posed problem and it admits steady state solutions (see Rognlien and Weinstock (1974)).

The coupled equations for the perturbed density and potential are given by (29) and (30). These equations are solved using a pseudo-spectral method code developed by Fyfe et al. (1977) and is based on the work of Orszag (1971). The dependent variables $\tilde{n}$ and $\tilde{\Phi}$ are Fourier-decomposed as follows:

$$\tilde{n}(\underset{\sim}{x},t) = \sum_{\underset{\sim}{k}} \tilde{n}(\underset{\sim}{k},t)\exp(i\underset{\sim}{k}\cdot\underset{\sim}{x}) \tag{47a}$$

$$\tilde{\Phi}(\underset{\sim}{x},t) = \sum_{\underset{\sim}{k}} \tilde{\Phi}(\underset{\sim}{k},t)\exp(i\underset{\sim}{k}\cdot\underset{\sim}{x}) \tag{47b}$$

where $\mathbf{k} = k_x \hat{e}_x + k_y \hat{e}_y$ and k_x and k_y are integers. The nonlinear $\mathbf{E} \times \mathbf{B}$ term (RHS of (29)) is computed by fast Fourier transforming $\nabla\tilde{n}$ and $\nabla\tilde{\Phi}$ from k space to x space, calculating the nonlinear term in x space by a simple multiplication, and then fast Fourier transforming the result back to k space. The equations are stepped forward in time in k space, and the cycle is repeated. Because of the imposition of conducting boundaries at $x = 0$ and $x = L$, we require that $v_x = 0$ at these boundaries (the plasma cannot flow through the boundaries). This leads to the following conditions on $\tilde{n}$ and $\tilde{\Phi}$: $\tilde{n}(k_x,k_y) = -\tilde{n}^*(k_x,k_y)$ and $\tilde{\Phi}(k_x,k_y) = -\tilde{\Phi}^*(k_x,k_y)$.

We present some representative results of this type of simulation study in Figs. 8 and 9. These results are from the paper by Hassam et al. (1986) in which the nonlinear evolution on the Rayleigh-Taylor instability was investigated. We mention that (29) and (30) are equivalent to the equations used in Hassam et al. (1986) if we make the identification that $V_0 = g/\Omega_i \nu_{in}$. The numerical results are computed on a 32 × 32 mesh; in order to prevent aliasing of the wave energy during the fast Fourier transformation, all modes with $|\mathbf{k}| > 32/3$ are set to zero amplitude. The plasma was initialized with small amplitude, randomly distributed modes. Two stages were clearly discernible in the temporal evolution of the system: (1) the linear stage in which the $\mathbf{k}$ spectrum develops peaks at small k_x and large k_y in accordance with linear theory (see (27)), and (2) the nonlinear stage in which the system reaches a turbulent state with a broad spectrum of modes.

In Fig. 8 we plot the total energy in the system as a function of time where the energy is defined as

$$E = \frac{1}{2} \sum_{\underset{\sim}{k}} k^2 |\phi_{\underset{\sim}{k}}|^2 \tag{48}$$

Initially there is a slight decrease in energy because many of the modes are damped. However, this is followed by the linear phase in which the unstable modes grow exponentially up until $t \sim 50$. Following this the plasma modes are strongly nonlinear and the instability saturates; the total energy is roughly constant for $t > 100$.

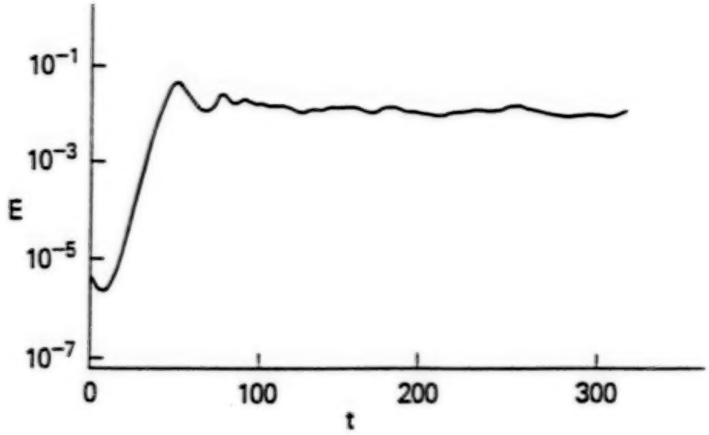


Fig. 8) Plot of total wave energy (E) vs time (t) during the evolution of the $\underset{\sim}{E} \times \underset{\sim}{B}$ instability (from Hassam et al. (1986)).

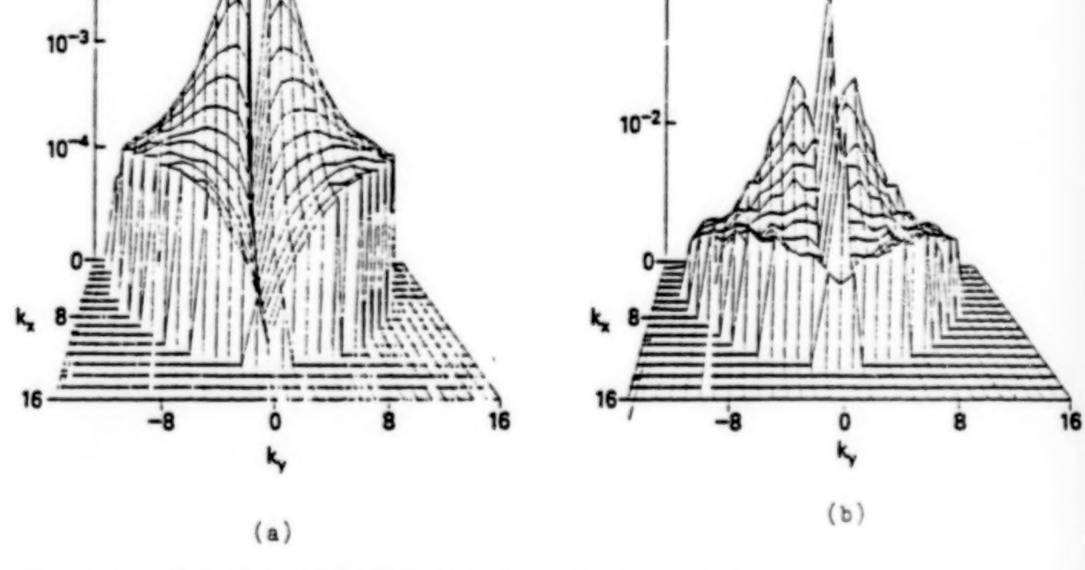


Fig. 9) Plot of the two dimensional, time averaged mode structure of the electrostatic potential and density fluctuations for the $\underset{\sim}{E} \times \underset{\sim}{B}$ instability (from Hassam et al. (1986)).

After the total energy saturates, the amplitudes of the wave modes can still be dynamic as energy is exchanged between different modes. The general characteristics of the spectra, however, remain unchanged: most of the energy resides in the longer wavelengths, and a power law type of falloff to shorter wavelengths is found. In order to see these features we present show Fig. 9 in which the time-averaged potential and density fluctuation spectra are shown in $\underset{\sim}{k}$ space. Note that the spectra have been reasonably smoothed out by the averaging process. Clearly, the bulk of the power resides in the long wavelength modes, small k_x and k_y. Interestingly, the potential spectrum is very anisotropic with relatively little energy in the region around $k_y \sim 0$. This is consistent with (30) in which the RHS is zero for $k_y = 0$. On the other hand, the density spectrum shows a peak in amplitude around the region $k_y \sim 0$. This is a reflection of the fact that while on average the turbulence tends to flatten the background density profile (i.e., $k_y = 0$), the fact that we hold the density fixed at $x = 0$ and $x = L$ results in sharp gradients near the boundaries leading to a substantial amount of power in the $k_y \sim 0$ density fluctuations.

Finally, we present the time-averaged power spectra (P_x and P_y) of the potential in Fig. 10. Here, P_x and P_y are defined as

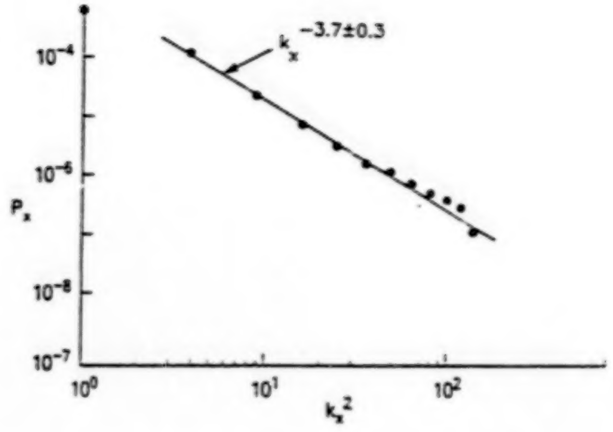
$$P_x(k_x) = \sum_{k_y} |\phi_{\underset{\sim}{k}}|^2 \tag{49a}$$

$$P_y(k_y) = \sum_{k_x} |\phi_{\underset{\sim}{k}}|^2 \tag{49b}$$

The approximate spectral indices, obtained by a least squares fit, are also shown in Fig. 10. We note that there is a definite anisotropy as indicated by the difference in the slopes of P_x and P_y; specifically, we find that $P_x \propto k_x^{-3.7\pm0.3}$ and $P_y \propto k_y^{-4.6\pm0.3}$.

Finally, we comment that the above discussion has not focussed on the quantitative details of the simulations. The purpose of this section is to give the reader a general understanding of the nonlinear, many-mode coupling process, an example of one simulation method used to study this process, and examples of the type of insights that can be gained from this study (e.g., wave spectra, spectral indices). We note that this type of analysis can also be carried using finite difference simulation codes and we refer the interested reader to Keskinen and Ossakov (1983).

(2) Large-Scale Structure. We now discuss nonlinear simulations of the $\underset{\sim}{E} \times \underset{\sim}{B}$ instability relevant to the regime $kL_n \sim O(1)$. A number of simulations in this regime has been carried out at the Naval Research Laboratory over the past decade [Zalesak and Ossakov, 1980; Keskinen and Ossakov, 1982; Zalesak et al., 1985; Mitchell et al., 1985; Huba et al., 1988a,b]. The first point we wish to stress is that the full nonlinear equations are solved numerically. That is, the dependent variables (density and potential) are solved exactly; they are not split into equilibrium and perturbation quantities. Thus, there is no need to impose the condition that fluctuating quantities remain small as in the previous section. Moreover, the ambient density profile can evolve dynamically and self-consistently with the changes in the electrostatic potential. For simplicity, we show results from a simulation in the collisional regime ($\nu_{in} \gg \omega$). This corresponds to the situation where the density and potential perturbations are confined to the F region, i.e., there is no coupling to the magnetosphere. We discuss a magnetosphere-ionosphere coupling model of



(a)

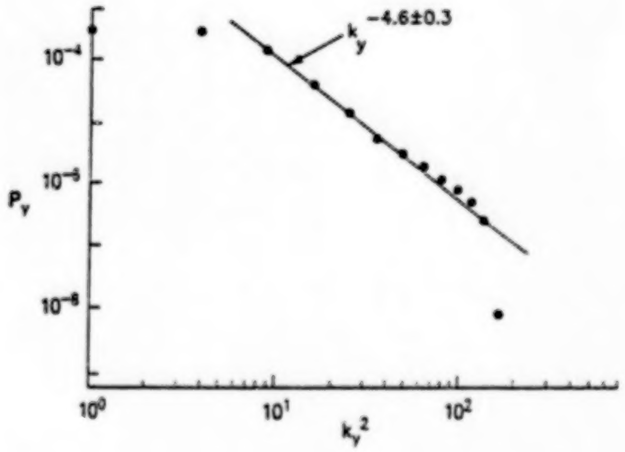


Fig. 10) Plot of the power spectral density [P_x and P_y] associated with the electrostatic potential ϕ (from Hassam et al. (1986)).

plasma instabilities in the Appendix. The equations being solved are continuity and current conservation:

$$\frac{\partial n}{\partial t} + \nabla \cdot n\underset{\sim}{V}_e = 0 \qquad (50)$$

$$\nabla \cdot n\nabla\phi = 0 \qquad (51)$$

where $\underset{\sim}{V}_e = -c\nabla\phi/B \times \hat{e}_z$.

The numerical methods used to solve (49) and (50) are described in Zalesak et al. (1985). The continuity equation (49) is solved using the multi-dimensional flux-corrected techniques of Zalesak (1979) while the potential equation (50) is solved with the incomplete Cholesky conjugate gradient algorithm of Hain (1980). The simulations presented here are performed on a 100 × 80 grid (x,y). The ambient density is initially characterized by a 1.5 to 1.0 density enhancement with a Gaussian profile of scale size 6 km in the x-direction and uniform in the y-direction. The ambient magnetic field is in the z-direction (B_0 = 0.5 G), and an ambient electric field is in the y-direction (E_0 = 0.025 V/m). The entire density enhancement $\underset{\sim}{E} \times \underset{\sim}{B}$ drifts in the x-direction with a velocity V_x = 500 m/sec. Periodic boundary conditions are assumed in the y-direction, and the grid is initialized with a random 1% density fluctuation.

The results of the simulation in shown in Fig. 11. The behavior of the density profile is typical of that observed in previous simulations [Keskinen and Ossakov, 1983]. In panel 2 (t = 520 sec) we see that a set of "fingers" has begun to form. The high density fingers grow outward into the low density background, while the low density fingers penetrate into the high density region. Subsequent nonlinear evolution involves the continued elongation of these fingers in the x-direction, with very little apparent change in their size in the y-direction. The original density enhancement is ultimately sliced into a group of sheets parallel to the initial density gradient.

Although Fig. 11 clearly shows the development of plasma structure in the plane transverse to B, one cannot directly compare Fig. 11 with any satellite or rocket data which is only one dimensional. Thus, in order to facilitate the comparison between the observational data and the simulation results, we plot density versus distance in Fig. 12 for three different satellite paths. We define θ to be the angle between the satellite path ($\underset{\sim}{x}_s$) and the -x-direction, i.e., $\theta = \cos^{-1}(-\underset{\sim}{x}_s \cdot e_x/|\underset{\sim}{x}_s|)$ (see Fig. 11). The striking features of Fig. 12 are the following. First, the perceived scale size of the ambient density profile varies considerably for the different paths. The actual scale size is roughly 6 km which is evident in the top panel (θ = 0°). However, for paths that slice the gradient at an angle one may be "fooled" into thinking that the scale size of the density profile is much larger than 6 km. Second, the density fluctuations look very different depending on the path. The top panel (θ = 0°) suggests that the density isn't very structured because the path only intersects one striation (or elongation). However, for paths which cut across many striations, as in the bottom two panels (θ = 63° and 83°), the density looks highly structured. Also, for the case θ = 83°, there is a considerable amount of plasma structure for x > 40 km, a region where there is no apparent background density gradient which could drive the instability. This is primarily a reflection of the fact that the instability has proceeded far enough to effectively destroy the original plasma structure.

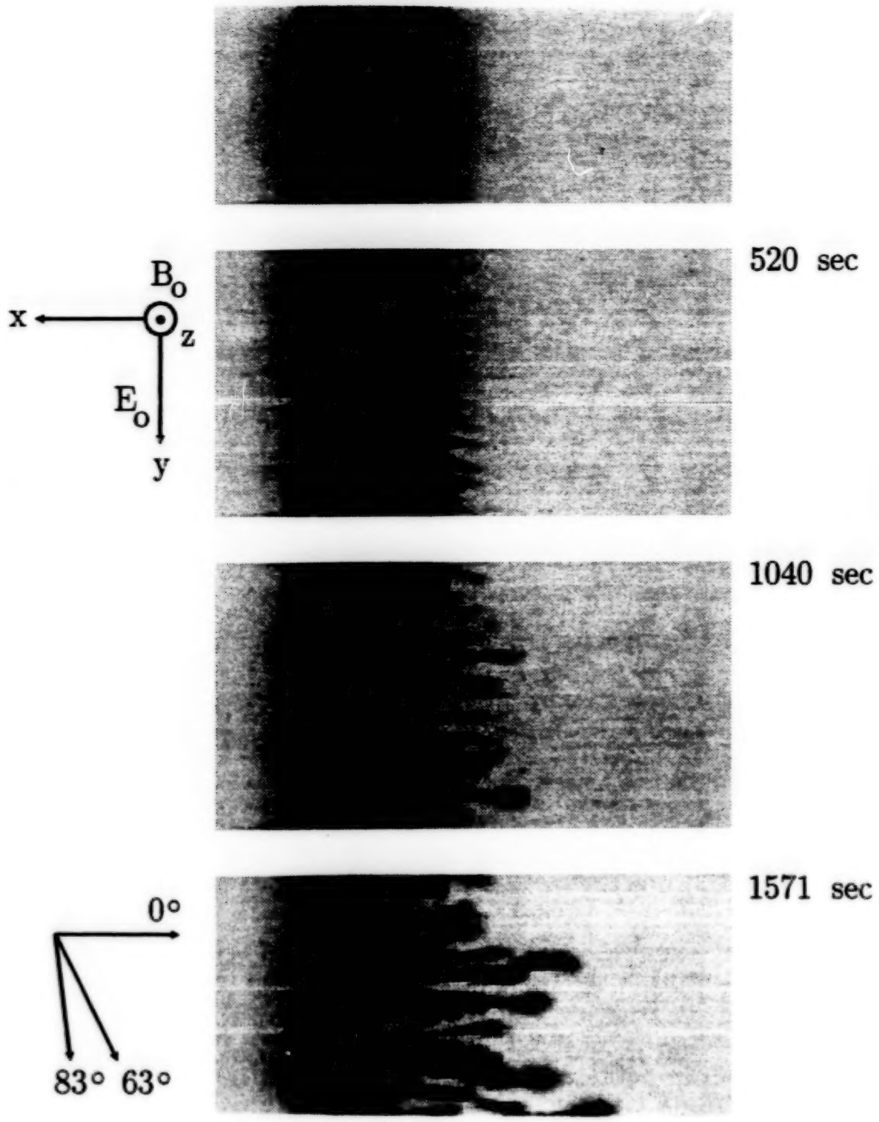


Fig. 11) Temporal evolution of the collisional **E** x **B** instability as it structures a Gaussian density enhancement.

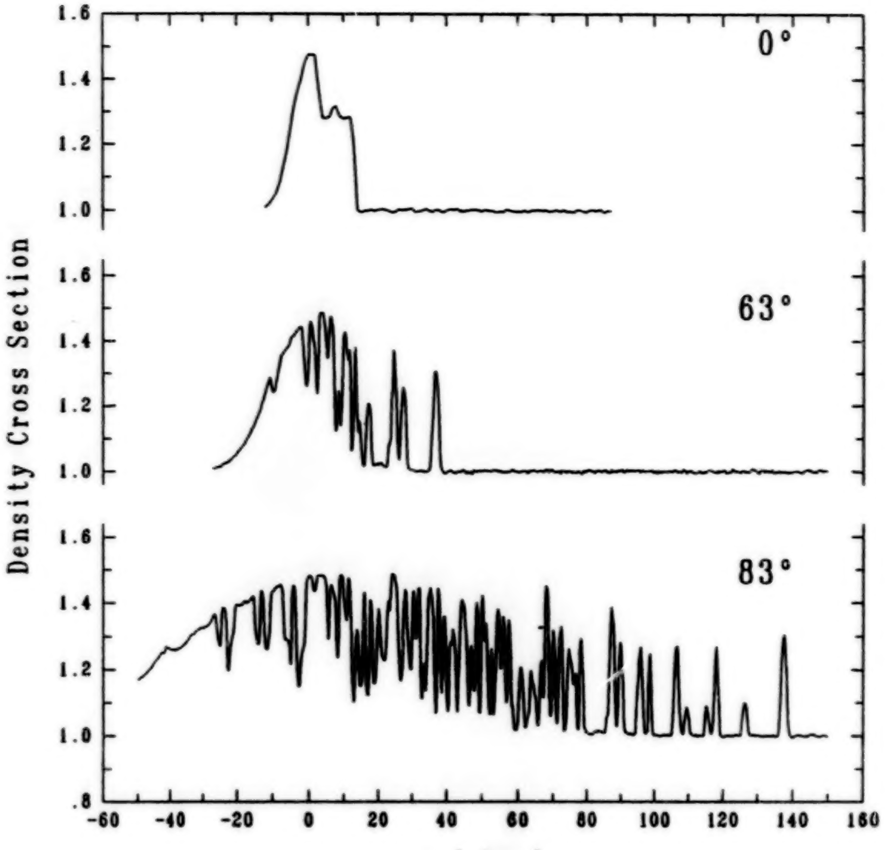


Fig. 12) Plot of the density (n) vs distance (x_s) where the path corresponds to a satellite moving at angle θ relative to the x-direction [(a) θ = 0°, (b) θ = 63°, (c) θ = 83°]. It is important to note how different the structure looks depending on the path.

Given this background, we now discuss experimental observations which support the contention that $\mathbf{E} \times \mathbf{B}$ instability is the source of high latitude turbulence (at times).

V. EXPERIMENTAL OBSERVATIONS

As noted by Tsunoda (1989) the most convincing evidence that the $\mathbf{E} \times \mathbf{B}$ instability is active in the high latitude ionosphere is reported by Cerisier et al. (1985) based on data from the Aureol 3 satellite. In this paper, electric field and density fluctuations, with reported scale sizes of 100's m to a few km, are found to occur asymmetrically on larger scale density enhancements (sizes of the order of 10's km). That is, on one side of the density enhancement plasma turbulence is observed, while on the other side the density profile is smooth. This is shown very clearly in Fig. 13 which is from Tsunoda (1989). Note in the top panel that the left hand sides of the density enhancements are structured, while the right hand sides, although much steeper, are unstructured. This is consistent with the necessary condition for the onset of the $\mathbf{E} \times \mathbf{B}$ instability which requires that $\mathbf{E}_0 \times \mathbf{e}_z \cdot \nabla n_0 > 0$. Cerisier et al. (1985) found that this condition was satisfied for more than half of the observations of the type shown in Fig. 13 (i.e., 4 out of 6). However, it should also be pointed out that it is very difficult for a single satellite to unambiguously determine whether or not this criterion is satisfied. This is because only the component of the density gradient along the path of the satellite is known. The real gradient (i.e., shortest density gradient scale length) can have any direction within 90° of the observed one.

To illustrate the difficulties of interpreting data from only a single satellite, we plot the density profiles based on our simulation results (shown in Fig. 12) as a function of time instead of space. That is, we will assume that a spacecraft passes through the plasma structure observed in the simulations at a velocity of ~ 7 km/sec. The results are shown in Fig. 14. Note that when θ is large, the "perceived" gradient scale length can be considerably larger than the actual scale length of 6 km. By the same token, the perceived scale size of the plasma structure also varies depending on the path as noted on Fig. 14. Thus, although it is difficult to make a definitive, quantitative comparison between observational data and simulation results, the evidence for the occurrence of the $\mathbf{E} \times \mathbf{B}$ instability in the high latitude ionosphere is very compelling.

VI. SUMMARY

We have presented a general overview of the theoretical and computational methods used to study plasma instabilities in the high latitude ionosphere. We illustrated the various methods with the $\mathbf{E} \times \mathbf{B}$ instability being the "guinea pig". The fundamental concepts and techniques are rather general and can be applied to other instabilities. Aside from going through several types of analytical and numerical calculations, we also tried to emphasize the physical conditions for which the different methods are valid, and the limitations of each method. It is also important to mention that, for pedagogical purposes, we have considered only the simplest model possible to describe the $\mathbf{E} \times \mathbf{B}$ instability in the high latitude ionosphere. There are a host of other physical processes which have been neglected, but should be included in a self-consistent manner. One major issue involves the coupling of the wave fields to the E-region and the magnetosphere, and understanding how the coupling impacts the evolution of the instability. Preliminary calculations suggest that coupling can

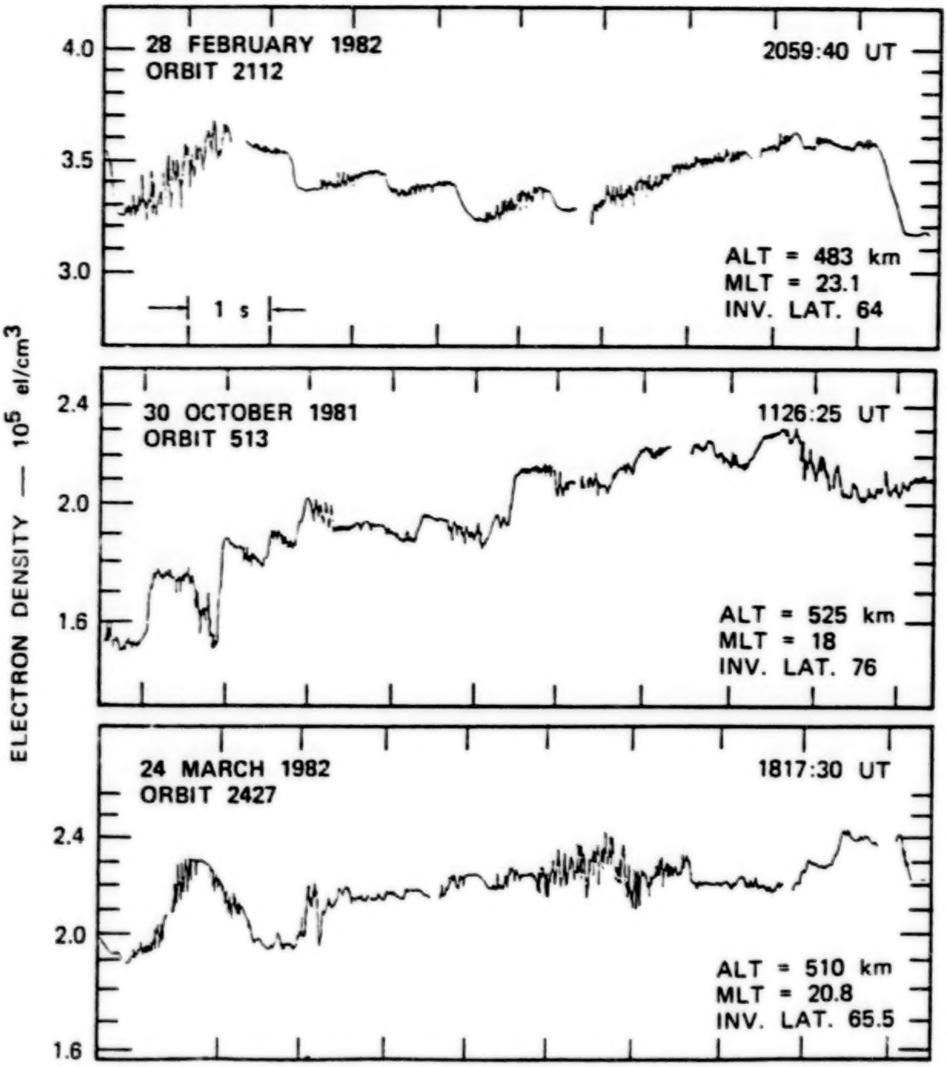


Fig. 13) Plot of density (n) vs time (t) from Aureol 3 satellite data (from Tsunoda (1989)).

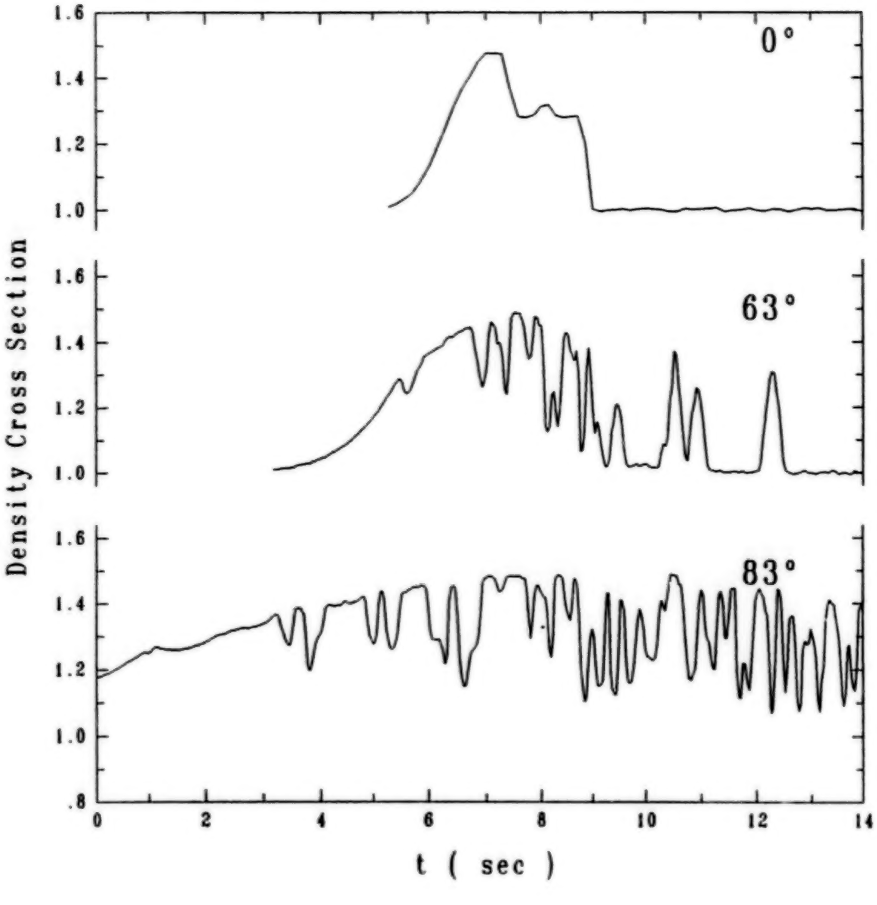


Fig. 14) Plot of density (n) vs time (t) based on the data in Fig. 12. We transform to from space to time assuming an orbital velocity V = 7 km/sec.

dramatically affect structure evolution and are discussed in the Appendix [Mitchell et al., 1985].

Closely related to the coupling issue is the issue of three dimensional dynamics, i.e., relaxing the equipotential field line approximation. This is a difficult problem and extremely pertinent to virtually all high latitude physical processes. Some research has been done regarding the linear theory of the $\mathbf{E} \times \mathbf{B}$ instability including parallel dynamics [Sperling et al., 1984; Drake et al., 1985; Huba and Chaturvedi, 1987]. Recently, there has been the successful development of a 3D electrostatic code to describe the evolution of ionospheric barium clouds which can be applied to high latitude processes [Drake et al., 1988; Zalesak et al., 1988].

Finally, it should be noted that the subject of high latitude ionospheric turbulence is very complex and not completely understood at this time. Although the $\mathbf{E} \times \mathbf{B}$ instability appears to be successful in explaining some observations in the sub-kilometer range, it certainly cannot explain all of them. In particular, some recent observations [Kintner, 1976; Basu et al., 1986; Basu et al., 1988] suggest that velocity shears are playing a role in the development of plasma structure. These observations have spurred interest in how velocity shear impacts the $\mathbf{E} \times \mathbf{B}$ instability [Huba et al., 1983], and in the Kelvin-Helmholtz instability [Kintner and Seyler, 1985; Keskinen et al., 1988] and other velocity shear instabilities [Ganguli and Palmadesso, 1988; Ganguli et al., 1988; Basu and Coppi, 1988].

ACKNOWLEDGMENTS

Discussions with my colleagues at the Naval Research Laboratory (M.J. Keskinen, J.A. Fedder, S.L. Ossakow, S.T. Zalesak, P.K. Chaturvedi) and Science Applications International Corporation (P. Satyanarayana, H.G. Mitchell, A.B. Hassam, J.F. Drake) over the past decade have been an enormous help to me in understanding instability processes in the ionosphere (and elsewhere), and are gratefully acknowledged. I especially thank H.G. Mitchell for providing the graphics and data analysis associated with the simulations discussed in Section IV.B.2 and the Appendix. This work has been supported by ONR and DNA.

APPENDIX: MAGNETOSPHERE-IONOSPHERE COUPLING MODEL

A relatively simple model of the effect of magnetosphere-ionosphere coupling on high latitude instabilities has been developed at NRL [Mitchell et al., 1985; Keskinen et al., 1988]. We have described this model in the first WITS Handbook [Huba et al., 1988b]. We include another discussion of the subject here for the sake of completeness, and to relate the theoretical analysis of Section III to this model.

We first discuss some underlying concepts associated with ionosphere-magnetosphere coupling and the evolution of plasma instabilities. The $\mathbf{E} \times \mathbf{B}$ instability is primarily an electrostatic flute mode; that is, it generates perturbations in the density and/or electric field transverse to the magnetic field $\mathbf{B}$ rather than parallel to $\mathbf{B}$. Moreover, the magnetic field remains constant, i.e., unperturbed. Alternatively, one can say that the perturbed transverse electric fields map perfectly along $\mathbf{B}$ (i.e., the magnetic field lines are equipotential lines). [Strictly speaking, it is expected that (1) the magnetic field lines are not perfect equipotential lines and (2) the magnetic field will be disturbed by any instability.] We note the following. First, the transverse electric field will have an e-folding distance along the magnetic field given by $L_{\|} \simeq k_{\perp}^{-1} (\sigma_{\|}/\sigma_{\perp})^{1/2}$ where $k_{\perp}$ is the transverse wave number of the electric field perturbation and $\sigma_{\|(\perp)}$ is the parallel (perpendicular) conductivity [Farley, 1959]. For the high latitude F region we note that $\sigma_{\|}/\sigma_{\perp} \gtrsim 10^6$; thus, kilometer scale irregularities can map well into the magnetosphere. Second, the disturbance will propagate along the magnetic field at the Alfvén velocity. The parallel distance it will travel is given by $L_A \simeq V_A/\gamma$ where γ is the characteristic growth rate of the instability and V_A is the Alfvén velocity. For the high latitude F region we note that $\gamma \lesssim 0.1$ sec and $V_A \gtrsim 10^3$ km/sec so that $L_A \gtrsim 10^4$ km. Again, on the time scale of the instability, the modes can map well into the magnetosphere. Conversely, modes generated in the magnetosphere can couple to the ionosphere.

Given this background we now describe our model. To simplify the analysis we will assume the modes are electrostatic. This means that the time scales of interest in the problem are long compared to the transit time of an Alfvén wave across the modeled region along the magnetic field. This assumption implies that the plasma is quasi-neutral everywhere, and that the electric field perpendicular to the magnetic field maps perfectly along the field within the modeled region. Further, since the electrons have a greater mobility parallel to the magnetic field than the ions, we assume that parallel currents are electron currents and that the ions have no velocity parallel to the field. As a result of these assumptions, the model may be viewed as a set of ion layers perpendicular to the magnetic field which are strongly linked by the mapping of the perpendicular electric field between layers and by the parallel electron currents which flow between layers in order to preserve quasi-neutrality. This is shown in Fig. 15.

For each layer, the governing equations are the continuity and momentum equations for each species, i.e., (1) and (2). In order to maintain quasi-neutrality (i.e., $\nabla \cdot \mathbf{J} = 0$) it is assumed that all currents generated within this model must also close within this model. The field-line integrated divergence of the perpendicular current must therefore be zero, i.e.,

$$\begin{aligned} 0 &= \int dz\ \nabla_\perp \cdot \underset{\sim}{j}_\perp = \nabla_\perp \cdot \left(\underset{\sim}{J}_{\perp \mathrm{Ped}} + \underset{\sim}{J}_{\perp \mathrm{pol}}\right) \\ &= \nabla \cdot \left(\Sigma_p\left[-\nabla_\perp \phi\right] - C_M\left[\frac{\partial}{\partial t} + \frac{c}{B}\left(\hat{e}_z \times \nabla_\perp \phi\right) \cdot \nabla\right]\nabla_\perp \phi\right) \end{aligned} \tag{A1}$$

where $\underset{\sim}{J}_{\perp \mathrm{Ped}}$ and $\underset{\sim}{J}_{\perp \mathrm{pol}}$ are the field-line integrated Pedersen and polarization currents (see Fig. 15), $\Sigma_p = \int(nec\nu_i/B\Omega_i)dz$ is the field-line integrated Pedersen conductivity, and $C_M = \int(nec/B\Omega_i)dz$ is the field-line integrated inertial capacitance. For details of this derivation see Mitchell et al. (1985). Thus, the plasma on each level E × B drifts under the influence of the perpendicular electric field ($\underset{\sim}{E} = -\nabla\phi$), which is itself determined by the fact that the perpendicular Pedersen and polarization drift currents driven by the field must close by means of parallel electron currents within the model.

The Pedersen currents tend to decay any potential which is not supported by the density gradients, while the polarization drift currents tend to slow this decay. As a result, the balance between these effects defines an inertial relaxation rate $\nu = \Sigma_p/C_M$ for changes in the potential. The relative importance of Pedersen versus polarization currents is dictated by the parameter ν/γ where γ is the linear growth rate of the instability under investigation. Pedersen currents dominate when $\nu/\gamma >> 1$ and polarization currents dominate when $\nu/\gamma << 1$. Thus, ion inertia is playing a significant role in the evolution of the instability when magnetospheric coupling is important; this is the reason we included ion inertia in the theoretical analysis in this paper.

To demonstrate the effect of magnetosphere-ionosphere coupling on the evolution of the $\underset{\sim}{E} \times \underset{\sim}{B}$ gradient drift instability we present the results of a simulation in which magnetospheric coupling is dominant. These results should be contrasted to those shown in Fig. 11 in which magnetospheric coupling is ignored. The physical configuration is the same as that described for Fig. 11. We only consider structure in the plane transverse to the ambient magnetic field, i.e., the xy plane. The F-layer is initially characterized by a 1.5 to 1.0 density enhancement with a Gaussian profile of scale size 12 km in the x-direction and uniform in the y-direction, a uniform magnetic field in the z-direction (B_z = 0.5 G), and a background electric field in the y-direction (E_y = .025 V/m). The entire enhancement E x B drifts in the x-direction at a velocity v_x = 0.5 km/sec. Finally, we consider $\nu/\gamma = 0.20$ so that the polarization current dominates over the Pedersen current. A uniform horizontal magnetosphere is assumed above the F-layer linked by the vertical magnetic field lines. The back edge of the F-layer enhancement, relative to the drift, is unstable to the $\underset{\sim}{E} \times \underset{\sim}{B}$ gradient drift instability. The simulations are performed on an 100 x 80 cell grid (x,y) with a cell size of 1.0 km x .25 km which is drifting with the enhancement at the E_y x B velocity. Periodic boundary conditions are assumed in the y-direction, and the grid is initialized with a random 1% density fluctuation.

The evolution of structure in the inertial regime (i.e., strong ionosphere-magnetosphere coupling) is shown in Fig. 16, and is very different from the collisional case (see Fig. 11). The linear growth rate for this case is $\gamma = 5.0 \times 10^{-3}$ sec^{-1} (or a growth time of τ = 200 sec) so that the growth of the mode is retarded by ionosphere-magnetosphere coupling. During the time period between panel 1 and panel 2 (t < 704 sec), the growth of the instability closely resembles that of Fig. 11. There is development of long, narrow density fingers. However, for t > 700 seconds,

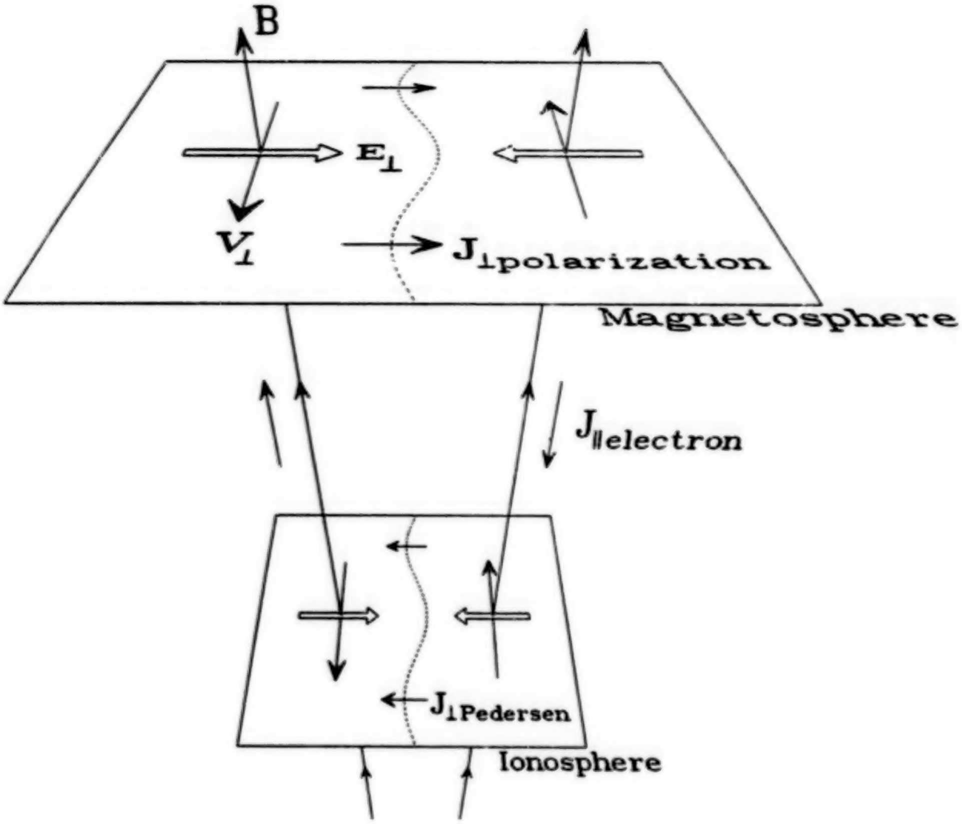


Fig. 15) Schematic of the physical model used for magnetospheric-ionospheric coupling. The coupling between the ionosphere and the magnetosphere is maintained by parallel electron currents flowing along equipotential field lines. The perpendicular currents are the Pedersen current in the ionosphere, and the polarization current in the magnetosphere.

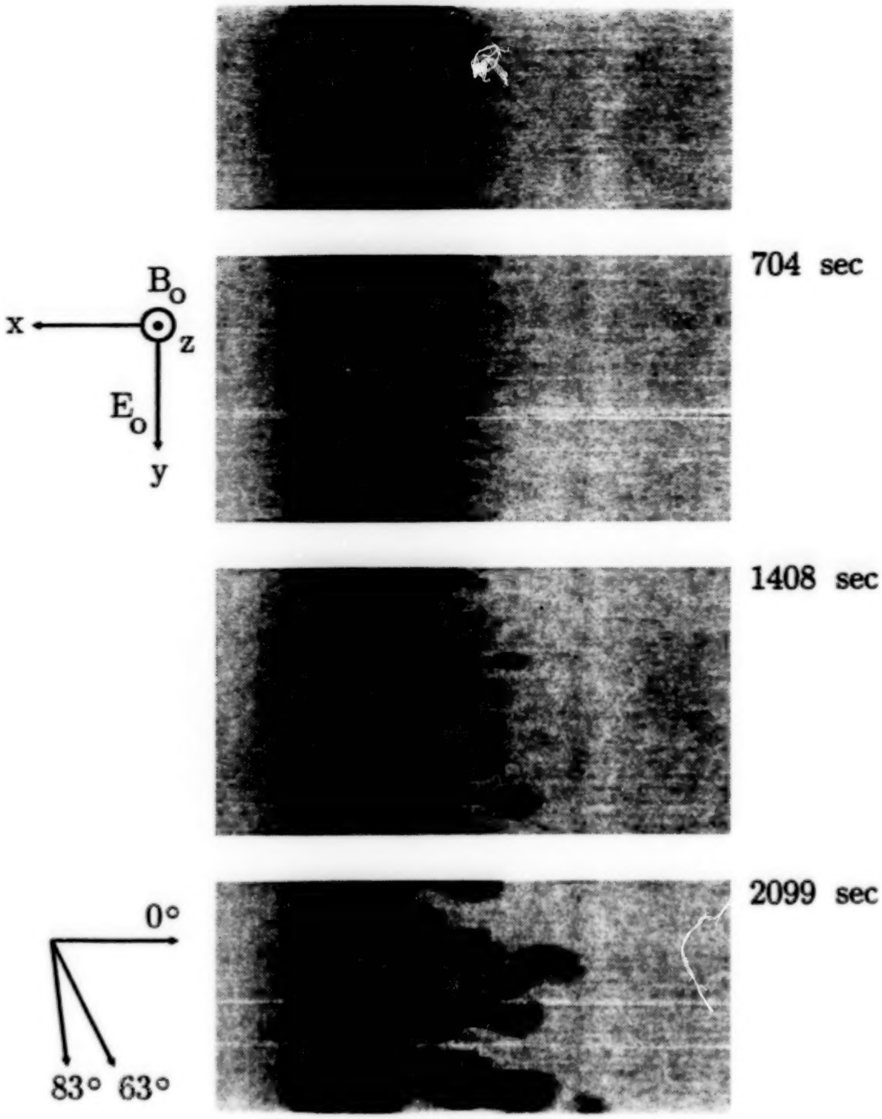


Fig. 16) Temporal evolution of the inertial $\mathbf{E} \times \mathbf{B}$ instability as it structures a Gaussian density enhancement. In this situation the magnetosphere is playing an important role in the structuring process.

the behavior changes radically from that in Fig. 11. In panels 3 and 4, we see that the fingers form mushroom-like heads and tend to thicken. No longer are they long, thin interpenetrating fingers; rather they are fat interpenetrating blobs. Any narrow fingers which begin to form quickly go to a mushroom shape and then spread out. In a number of simulations we have noted a tendency for the structure in the y direction to undergo an inverse cascade to the longest mode which will fit in the system. This feature can clearly be seen in panel 4, where the structured state throughout most of the simulation region shows two blobs: one of high density, and the other of low density.

REFERENCES

Baker, K.B., R.A. Greenwald, and R.T. Tsunoda, Very high latitude F-region irregularities observed by HF-radar backscatter, Geophys. Res. Lett., 10, 904, 1983.

Basu, B. and B. Coppi, Fluctuations associated with sheared velocity regions near auroral arcs, Geophys. Res. Lett., 15, 417, 1988.

Basu, Su, S. Basu, C. Senior, D. Weimer, E. Nielsen, and P.F. Fougere, Velocity shears and sub-km scale irregularities in the nighttime auroral F-region, Geophys. Res. Lett., 13, 101, 1986.

Basu, Su., S. Basu, E. MacKenzie, P.F. Fougere, W.R. Coley, N.C. Maynard, J.D. Winningham, M. Sugiura, W.B. Hanson, and W.R. Hoegy, Simultaneous density and electric field fluctuation spectra associated with velocity shears in the auroral oval, J. Geophys. Res., 93, 115-136, 1988.

Cerisier, J.C., J.J. Berthelier, and C. Beghin, Unstable density gradients in the high latitude ionosphere, Radio Sci., 20, 755, 1985.

Chandrasekhar, S., Hydrodynamic and Hydromagnetic Stability, (Clarendon, Oxford, 1961).

Chaturvedi, P.K., and S.L. Ossakow, Nonlinear theory of the collisional Rayleigh-Taylor instability in equatorial Spread F, Geophys. Res. Lett., 4, 558, 1977.

Chaturvedi, P.K., and S.L. Ossakow, Nonlinear stabilization of the $\mathbf{E} \times \mathbf{B}$ gradient drift instability in ionospheric plasma clouds, J. Geophys. Res., 84, 419, 1979.

Chaturvedi, P.K., and S.L. Ossakow, The current convective instability as applied to the auroral ionosphere, J. Geophys. Res., 86, 4811, 1981.

Chaturvedi, P.K. and J.D. Huba, The interchange instability in high latitude plasma blobs, J. Geophys. Res., 92, 3357, 1987.

Drake, J.F., J.D. Huba, and S.T. Zalesak, Finite temperature stabilization of the gradient drift instability in barium clouds, J. Geophys. Res., 90, 5227, 1985.

Drake, J.F., M. Mulbrandon, and J.D. Huba, Three-dimensional equilibrium and stability of ionospheric plasma clouds, Phys. Fluids, 31, 3412, 1988.

Farley, D.T., A theory of electrostatic fields in a horizontally stratified ionosphere subject to a vertical magnetic field, J. Geophys. Res., 64, 1225, 1959.

Fyfe, D., G. Joyce, and D. Montgomery, Magnetic dynamo action in two-dimensional turbulent magnetohydrodynamics, J. Plasma Phys., 17, 317, 1977.

Ganguli, G. and P.J. Palmadesso, Electrostatic ion instabilities in the presence of parallel currents and transverse electric fields, Geophys. Res. Lett., 15, 103, 1988.

Ganguli, G., Y.C. Lee, and P.J. Palmadesso, Kinetic theory for electrostatic waves due to transverse velocity shears, Phys. Fluids, 31, 823, 1988.

Gary, S.P. and M. Thomsen, Collisionless electrostatic interchange instabilities, J. Plasma Phys., 28, 551, 1982.

Gary, S.P. and T.E. Cole, Pedersen density drift instabilities, J. Geophys. Res., 88, 10104, 1983.

Hain, K., A non-recursive incomplete Cholesky decomposition method for the solution of linear equations with a sparse matrix, NRL Memorandum 4264, Naval Research Laboratory, Washington, DC, (1980).

Hassam, A.B., W. Hall, J.D. Huba, and M.J. Keskinen, Spectral characteristics of interchange turbulence, J. Geophys. Res., 91, 13513, 1986.

Huba, J.D. and S.T. Zalesak, Long-wavelength limit of the $\underline{E} \times \underline{B}$ instability, J. Geophys. Res., 88, 10263, 1983.

Huba, J.D., S.L. Ossakov, P. Satyanarayana, and P.N. Guzdar, Linear theory of the $\underline{E} \times \underline{B}$ instability with an inhomogeneous electric field, J. Geophys. Res., 88, 425, 1983.

Huba, J.D., A.B. Hassam, I.B. Schwartz, and M.J. Keskinen, Ionospheric turbulence: Interchange instabilities and chaotic fluid behavior, Geophys. Res. Lett., 12, 65, 1985.

Huba, J.D., H.G. Mitchell, M.J. Keskinen, J.A. Fedder, P. Satyanarayana, and S.T. Zalesak, Simulations of plasma structure evolution in the high-latitude ionosphere, Radio Sci., 23, 503, 1988a.

Huba, J.D., H.G. Mitchell, M.J. Keskinen, J.A. Fedder, P. Satyanarayana, and S.T. Zalesak, Simulations of plasma structure evolution in the high-latitude ionosphere, WITS Handbook Vol. 1, ed. C.H. Liu and B. Edwards, 140, 1988b.

Huba, J.D., Theoretical and simulation methods applied to high latitude ionospheric turbulence, to appear in Proc. of MIT Space Sci. Sym., ed. T. Chang, 1989.

Huba, J.D., G. Ganguli, P. Satyanarayana, A.B. Hassam, and D. Winske, Generalized interchange instability in a finite β, collisional plasma, to be submitted to Phys. Fluids, 1989.

Kelley, M.C., J.F. Vickrey, C.W. Carlson, and R. Torbert, On the origin and spatial extent of high-latitude F region irregularities, J. Geophys. Res., 87, 4469, 1982.

Kelley, M.C., C.E. Seyler, and S. Zargham, Collisional interchange instability, 2, A comparison of the numerical simulations with the in situ experimental data, J. Geophys. Res., 92, 10089, 1987.

Keskinen, M.J. and S.L. Ossakov, On the spatial power spectrum of the $\underline{E} \times \underline{B}$ gradient drift instability in ionospheric plasma clouds, J. Geophys. Res., 86, 6947, 1981.

Keskinen, M.J. and S.L. Ossakov, Nonlinear evolution of plasma enhancements in the auroral ionosphere, 1, Long wavelength irregularities, J. Geophys. Res., 89, 144, 1982.

Keskinen, M.J. and S.L. Ossakov, Nonlinear evolution of convecting plasma enhancements in the auroral ionosphere, 2, Small scale irregularities, J. Geophys. Res., 88, 474, 1983.

Keskinen, M.J. and S.L. Ossakov, Theories of high-latitude ionospheric irregularities: a review, Radio Sci., 18, 1077, 1983.

Keskinen, M.J., H.G. Mitchell, J.A. Fedder, P. Satyanarayana, S.T. Zalesak, and J.D. Huba, Nonlinear evolution of the Kelvin-Helmholtz instability in the high-latitude ionosphere, J. Geophys. Res., 93, 137, 1988.

Keskinen, M.J., Scaling laws for the spectrum of interchange instabilities in the high latitude ionosphere, to be submitted to J. Geophys. Res., 1989.

Kintner, P.M., Observations of velocity shear driven plasma turbulence, J. Geophys. Res., 81, 5114, 1976.

Kintner, P.M. and C.E. Seyler, The status of observations and theory of high latitude ionospheric and magnetospheric plasma turbulence, Space Sci. Rev., 41, 91, 1985.

Krall, N.A. and A.W. Trivelpiece, Principles of Plasma Physics, (McGraw-Hill, New York, 1973).

Lorenz, E.N., Deterministic nonperiodic flow, J. Atmos. Sci., 20, 130, 1963.

Mitchell, H.G., J.A. Fedder, M.J. Keskinen, and S.T. Zalesak, A simulation of high latitude F layer instabilities in the presence of magnetosphere-ionosphere coupling, Geophys. Res. Lett., 12, 283, 1985.

Orszag, S.A., Numerical simulation of incompressible flows within simple boundaries. 1. Galerkin (spectral) representations, Stud. Appl. Math., 50, 293, 1971.

Ossakow, S.L., P.K. Chaturvedi, and J.B. Workman, High-altitude of the gradient drift instability, J. Geophys. Res., 83, 2691, 1978.

Rino, C.L., R.C. Livingston, and S.J. Matthews, Evidence for sheet-like auroral ionospheric irregularities, Geophys. Res. Lett., 5, 1038, 1978.

Roberts, K. and J.B. Taylor, Magnetohydrodynamic equations for finite Larmor radius, Phys. Rev. Lett., 8, 197, 1962.

Rodriquez, P. and E. Szuszczewicz, High latitude irregularities in the lower F-region: intensity and scale size distribution, J. Geophys. Res., 89, 5575, 1984.

Rognlien, T.D. and J. Weinstock, Theory of the nonlinear spectrum of the gradient drift instability in the equatorial electrojet, J. Geophys. Res., 79, 4733, 1974.

Sperling, J.L. and A.J. Glassman, Short-circuiting, ion-viscous, and ion-inertial effects in striation "freezing", J. Geophys. Res., 90, 8707, 1985.

Sperling, J.L., J.F. Drake, S.T. Zalesak, and J.D. Huba, The role of finite parallel length on the stability of barium clouds, J. Geophys. Res., 89, 10913, 1984.

Tsunoda, R.T., High-latitude F-region irregularities: A review and synthesis, Rev. of Geophys. and Spa. Sci., 1989 (in press).

Vickrey, J.F., C.L. Rino, and T.A. Potemra, Chatanika/Triad observations of unstable ionization enhancements in the auroral F region, Geophys. Res. Lett., 7, 789, 1980.

Zalesak, S.T., Fully multidimensional flux-corrected transport algorithms for fluids, J. Comp. Phys., 31, 335, 1979.

Zalesak, S.T. and S.L. Ossakow, Nonlinear equatorial spread F: Spatially large bubbles resulting from large horizontal scale initial perturbations, J. Geophys. Res., 85, 2131, 1980.

Zalesak, S.T., P.K. Chaturvedi, S.L. Ossakow, and J.A. Fedder, Finite temperature effects on the evolution on ionospheric barium clouds in the presence of a conducting background ionosphere I. A high altitude incompressible background ionosphere, J. Geophys. Res., 90, 4299, 1985.

Zalesak, S.T., J.F. Drake, and J.D. Huba, Dynamics of three-dimensional ionospheric plasma clouds, Radio Sci., 23, 591, 1988.

Zargham, S. and C.E. Seyler, Collisional interchange instability, 1, Numerical simulations of intermediate-scale irregularities, J. Geophys. Res., 92, 10073, 1987.

Chapter 16

THE DERIVATION OF MERIDIONAL NEUTRAL WINDS IN THE THERMOSPHERE FROM *F2*-LAYER HEIGHT

K. L. Miller

Center for Atmospheric and Space Sciences
Utah State University
Logan, UT 84322-4405

P. G. Richards and D. G. Torr

Center for Space Plasma and Aeronomic Research
The University of Alabama in Huntsville
Huntsville, AL 35899

ABSTRACT

The measurement of neutral air motions is an important part of the World Ionosphere-Thermosphere Study. This paper describes a recently developed method of deriving the meridional component of the horizontal neutral wind in the thermosphere. The method is based on the approximately linear relationship between the height of the ionospheric *F2*-layer peak and the strength of the meridional neutral wind. An ionospheric photochemical model is used to compute the parameters that describe this relationship. The neutral wind speed is derived from a measurement of the height of the *F2*-layer (h_{max}). The measurement of h_{max} can be made by incoherent scatter radar and ionogram inversion techniques, but is also derivable from parameters that are routinely scaled from ionograms and archived in world data centers. The wind speeds derived from h_{max} are compared with winds from Fabry-Perot interferometry and incoherent scatter radar, and with results of the NCAR Thermospheric General Circulation Model.

1. INTRODUCTION

The understanding of the coupling between the neutral winds of the thermosphere and the ionosphere is one of the goals of the WITS program. While the measurement of many of the properties of the ionosphere can be made reliably using radio propagation and radar techniques, the measurement of the properties of the neutral thermosphere is difficult at best.

A measurement of the neutral air motions is valuable in any study of ionosphere-thermosphere coupling. The neutral wind affects many of the observable quantities and physical processes of the ionosphere, including the density profile of the ionospheric *F* region, and the generation and maintenance of electric fields.

There have been several models developed to describe the photochemistry and dynamics of the thermosphere and ionosphere [Torr et al, 1988; Sojka, 1989]. The models are becoming more and more complex, and are to a great extent limited by the accuracies of input parameters such as reaction rates, collision cross-sections, and solar flux. One of the major limitations, especially to the electron density profile, is the lack of knowledge of the meridional neutral wind.

This paper describes a method to derive the meridional component of the neutral wind in the thermosphere from existing ionosonde measurements. The method has been shown to derive winds with comparable accuracy to other techniques. It has the advantage of being able to derive winds at any local time, and at any of the midlatitude ionosonde sites. Winds can be derived retroactively, since data from many ionosondes are routinely archived at the world data centers.

2. DEPENDENCE OF THE *F2*-LAYER HEIGHT ON MERIDIONAL WIND SPEED

Because of the magnetic control of the processes that determine the *F2*-layer height, the most useful coordinate system is one that reflects the direction of the geomagnetic field. Hence the use of the words "north" and "northward" in this paper always refer to magnetic north, and "meridional" always refers to the magnetic meridian. The declination of the magnetic field is generally less than 20° throughout the mid-latitude region. The largest declinations at mid-latitudes occur in the South Indian Ocean.

The component of the neutral wind that is parallel to the geomagnetic field affects the profile shape and thus the concentration of the ionization at *F*-region heights through ion-neutral collisions [Rishbeth, 1972]. Since vertical neutral drifts are normally small, the dominant collisional forcing is from the magnetically north-south component of the horizontal wind. A pole-ward component of the horizontal neutral wind will have the effect of forcing ionization downward parallel to the magnetic field, while an equatorward component will have the opposite effect.

The motion of the ions maintained by a meridional neutral wind is dependent on the dip angle, I, of the geomagnetic field. Under quasi-steady-state conditions, the contribution to the northward parallel ion velocity (V_i) from the magnetically-northward neutral wind component (U_N) is

$$V_i = U_N \cos(I) \tag{1}$$

The vertical component of the ion velocity (V_Z) that is due to this effect is

$$V_Z = -U_N \sin(I) \cos(I) \tag{2}$$

If there is no externally applied forcing of vertical motion, the shape of the *F2* layer is determined by production, loss, and diffusion. The motion of the ions that is induced through the motion of the neutral wind either adds to or subtracts from the diffusion velocity of the ions. The height of the maximum ion density, or balance height, is determined by the altitude where the influences of loss and diffusion are approximately equal. If a force with a component parallel to the magnetic field is applied by an external source such as neutral wind or electric fields, the peak electron density is displaced from the balance height.

The variation of sin(I)cos(I) with magnetic latitude is shown in Figure 1. The magnetic dip angle in this figure was computed using the IGRF magnetic field model [Peddie, 1982]. The calculation was done at 255° E longitude since this meridian in the IGRF model includes the geomagnetic pole. Figure 1 shows that the effectiveness of the neutral wind to affect the height of the *F2* layer decreases significantly at high and low latitudes. It is most effective near 25° from the geomagnetic dip equator where the dip angle is 45°. It is, however, a significant factor in determining the vertical ion motions in the *F* region throughout the midlatitude region.

3. DEVELOPMENT OF THE "SERVO" MODEL

Rishbeth [1966] and Hanson and Patterson [1964] showed that if the external forcing from the neutral atmosphere is not too great, there is a linear relationship between the height of the layer (h_{max}) and the neutral wind speed. Rishbeth [1967] compared the behavior of the F2 region under the influence of a meridional neutral wind to a servo system, and developed the equations describing the rate of vertical movement of the layer and the equilibrium position of the peak. According to the servo model, an equatorward wind will force the ionization to higher altitudes. The raising of the layer increases the downward diffusion, which in turn opposes further upward motion of the layer. Similarly, a poleward wind will force the ionization to lower altitudes where the increased recombination rate opposes a further lowering of the layer maximum. Except for extreme wind speeds, and if electric fields are small, a linear relationship holds approximately between the neutral flow (U_N) and the resulting change in the height of the *F*-layer peak [Rishbeth and Barron, 1960; Hanson and Patterson, 1964; Rishbeth, 1966; Buonsanto, et al. 1989].

By assuming an *F2* region composed primarily of O^+ with an ionization peak well above the production region and a loss process dominated by the reaction of O^+ with N_2, Rishbeth et al. [1978] showed that an externally induced vertical drift and the rate of change of the layer height are related approximately by

$$\frac{dh_m}{dt} = \frac{D_m \sin^2(I)}{2H}\{\exp[-k(h_m-h_o)/H]-\exp[(h_m-h_o)/H]\}+V_z \quad (3)$$

where

D_m = diffusion coefficient for O^+

I = magnetic dip angle

h_m = height of the *F2*-layer peak (h_{max})

h_o = balance height

H = scale height of the neutral ionizable gas

kH = scale height of O^+, which controls diffusion

V_z = applied vertical ion drift

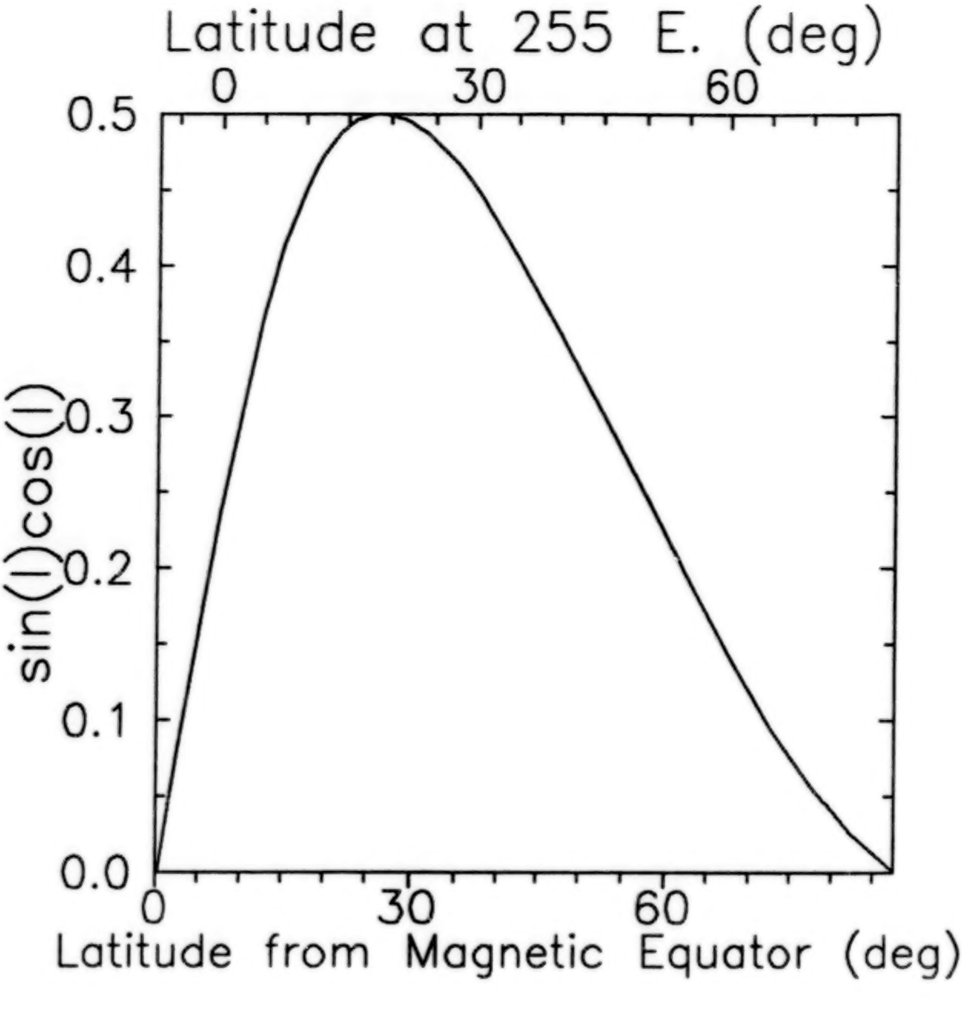


Figure 1. The ratio of vertical ion drift to horizontal neutral wind speed (sin(I) cos(I)) as a function of latitude north of the geomagnetic equator.

The plasma diffusion coefficient, D_m, is determined at the balance height, and is related approximately to the recombination coefficient at the peak, β_m, by

$$\beta_m = 0.628 \frac{D_m \sin^2(I)}{H^2} \tag{4}$$

in the daytime, and

$$\beta_m = 0.115 \frac{D_m \sin^2(I)}{H^2} \tag{5}$$

at night [Buonsanto, et al. 1989].

Equation 3 can be solved for the applied vertical drift, giving

$$V_Z = \frac{dh_m}{dt} + \{\exp[(h_m-h_o)/H]-\exp[-k(h_m-h_o)/H]\} \frac{D_m \sin^2(I)}{2H} \tag{6}$$

Buonsanto et al. [1989] showed that, for most conditions in the *F* region, the effects of both the time rate of change of the layer height in the first term of Equation 6, and the non-linearity in the second term are small. Under these conditions, the vertical drift becomes

$$V_Z = \frac{(k+1)\ D_m \sin^2(I)}{2H^2} (h_m-h_o) \tag{7}$$

If the vertical drift of the neutral atmosphere is assumed to be small, the vertical drift of the ionization can be assumed to be the result of a combination of ion collisions with the meridional component of the horizontal neutral wind and ion drift induced by the east-west electric field. Combining the E x B drift with Equation 2, the vertical drift can be written in terms of the northward neutral wind (U_N) and the eastward electric field (E_E) as

$$V_Z = -U_N \sin(I) \cos(I) + \frac{E_E \cos(I)}{B} \tag{8}$$

For conditions existing most of the time at mid-latitudes, the "servo" equation yields meridional neutral wind speeds from measurements of the *F2*-layer height and a neutral atmosphere model using a simple relationship from Equations 7 and 8:

$$U_N = \frac{(h_o - h_m)}{\alpha} + \frac{E_E}{B \sin(I)} \tag{9}$$

where

$$\alpha = \frac{2H^2 \cos(I)}{(k+1)D_m \sin(I)}$$

4. DEVELOPMENT OF THE "WIND FROM H_{MAX}" MODEL

Ionospheric models are dependent on winds and electric fields to define the profile shape and density of the *F* region. They contain within them the complex photochemistry that defines how h_{max} depends on the meridional wind speed. The models use this information in computing species concentrations when the neutral wind is specified.

If the east-west electric field is small, or if the effects of the meridional neutral wind and electric field are to be considered together, Equation 9 becomes a simple proportionality between wind and layer height:

$$U_N = \frac{(h_o - h_m)}{\alpha} \tag{10}$$

Miller et al. [1986] took advantage of the capability of the FLIP model [Richards and Torr, 1985, 1988; Young, et al., 1980ab, Chandler et al., 1983] to compute h_{max} for a specified wind speed and developed a method to derive an approximate constant of proportionality for the relationship between *F2*-layer height and meridional wind speed.

4.1. Derivation of α and h_o.

A plot of h_{max} vs. meridional neutral wind speed has the characteristic "S"-shape shown in Figure 2. Figure 2 also shows the day to night difference in the response of h_{max}. The greater electron density in the daytime, combined with larger production and recombination rates result in a much stronger response of the layer height to small wind speeds, and then a much more rapid end to the linear portion of the curve. On the other hand, the nighttime change in layer height is smaller, with the linear region extending to greater equatorward wind speeds. The day-to-night difference in the response of the layer height to the neutral wind is important to the accuracy of wind speeds derived using this technique. The wind during the daytime is normally pole-ward and relatively small, while the nighttime wind is equator-ward and often near 200 m s^{-1}.

The form of Equation 10 suggests that in the linear region the neutral wind speed could be derived directly from measurements of the height of the *F2*-layer peak if α and h_o are known. These parameters can be derived using an ionosphere model to derive layer heights at several wind speeds, including h_o at $U_N = 0$, and then determining the slope (α) of the linear region. In practice, little accuracy is lost by finding h_{max} at two wind speeds that span the expected wind speeds but are still within the linear region, and then determining α from the slope of the line connecting the two points and h_o from the height where the line crosses $U_N = 0$.

4.2. Measurement of h_{max}

One of the advantages of having the capability to derive meridional wind speeds from h_{max} is the availability of data on a global scale from which h_{max} can be derived. The more traditional optical and incoherent scatter radar techniques of measuring neutral winds are valuable and provide checks on this method, but are limited in their coverage

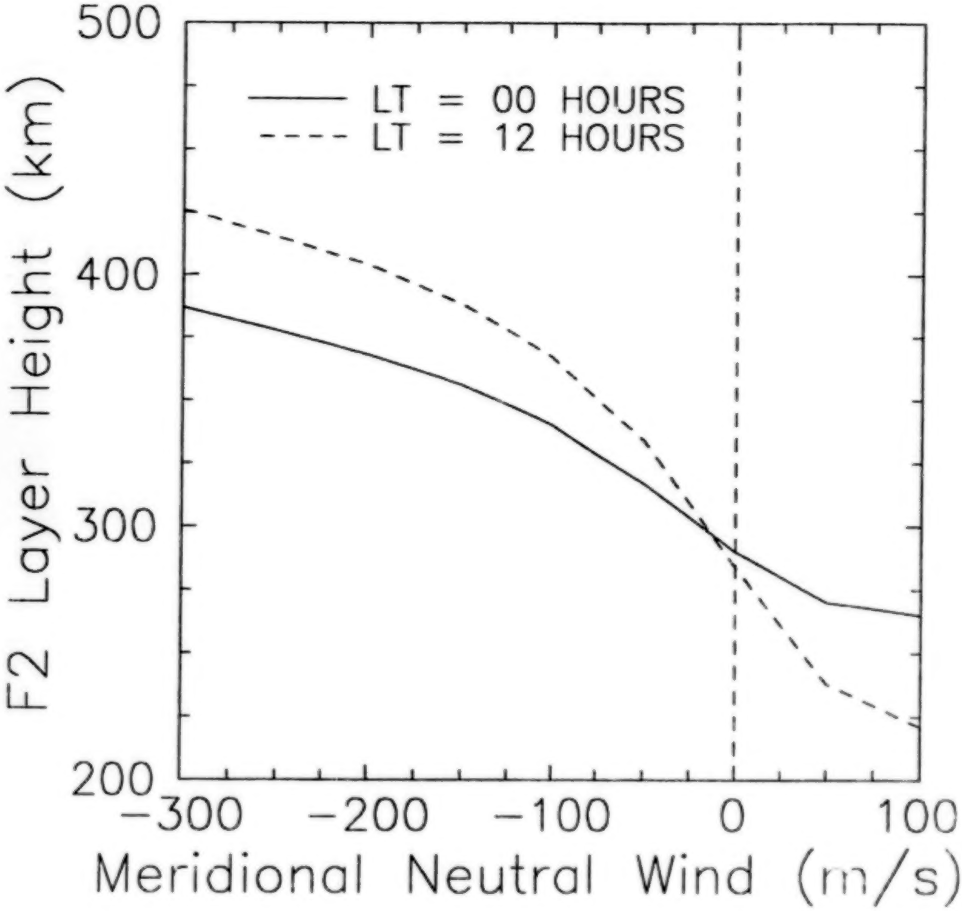


Figure 2. Height of the peak of the *F2* layer as a function of meridional wind speed at noon and midnight for equinox conditions.

by the weather and hours of darkness in the case of the optical measurements, and by the limited number and locations of the radars.

Ionosondes provide the data base from which global arrays of h_{max} can be derived. In simple terms, an ionosonde measures the time of flight of a radio pulse from the transmitter to the reflection point in the ionosphere and back to the ionosonde receiver. As the frequency increases, a virtual-height profile of the plasma frequency of the ionosphere, or ionogram, is generated. There is no reflection above the peak plasma frequency of the *F2* region *(foF2)*. To determine h_{max} the virtual height profile must be converted to a true height profile, or at least the true height of the reflection at *foF2* must be determined.

In the best situation, the ionogram is either recorded digitally or hand scaled, and a true-height analysis converts the vertical heights of reflection to true heights [Titheridge, 1985]. It is tedious work to hand-scale a complete ionogram, and is not done routinely. Complete electron density profiles are being recorded, however, by a new generation of digital HF sounders.

At the present, any global study that relies on ionosonde data must derive its parameters from data that are routinely scaled from ionograms and archived in the World Data Centers. The parameters most relevant to this study are the maximum plasma frequency of the *F2* region, *foF2*, the maximum frequency of the *E* region, *foE*, and the transmission factor *M(3000)F2*.

Dudeney [1983] reviewed the development of an empirical model that determines the layer height primarily from the approximately inverse relationship between the height of the peak of the *F2* layer, *hmF2* or h_{max}, and *M(3000)F2*. This model, based on work by Shimazaki [1955] and by Bradley and Dudeney [1973] uses *M(3000)F2* for the primary calculation of *F2* layer height, and uses the ratio of *foF2* to *foE* as an indicator of the underlying ionization that retards the pulse and increases the virtual height. According to Dudeney [1983] the most accurate representation of the model is written as

$$h_{max} = \frac{1490\ F}{M+\Delta M} - 176 \qquad (11)$$

where

$$F = M \left(\frac{0.0196\ M^2 + 1}{1.2967\ M^2 - 1} \right)^{1/2}$$

$$\Delta M = \frac{0.253}{X_E - 1.215} - 0.012$$

$$X_E = \frac{foF2}{foE}$$

$$M = M(3000)F2$$

4.3. Accuracy of Derived Winds.

It is difficult to assign a realistic error bar to wind speeds derived from the *F2*-layer height. The major difficulty lies in the fact that it is based on an ionosphere model that predicts average behavior. Although the FLIP model generates the ionization density from photochemical reactions, it uses the MSIS empirical model for neutral densities and neutral temperature. The model performs well in tests against measurements [Chandler et al., 1983], but is unable to predict unusual or transient features.

A statistical error of the uncertainty in U can be estimated from Equation 10:

$$\frac{\Delta U}{U} = \left[\left(\frac{\Delta h_m}{h_o - h_m}\right)^2 + \left(\frac{\Delta h_o}{h_o - h_m}\right)^2 + \left(\frac{\Delta \alpha}{\alpha}\right)^2\right]^{1/2} \tag{12}$$

Determination of Δh_o and $\Delta\alpha$ are based on the accuracy of the FLIP model. The height of the *F2*-layer peak is determined by a parabolic fit to the three points that define the peak ion density in the model. The step size can be adjusted, but is nominally about 8 km in the region of interest. A larger source of uncertainty is the degree to which the model represents the real ionosphere.

The value of h_o is derived from the point where the slope between h_m values at two wind speeds intersects U=0:

$$h_o = \frac{U_2 h_1 - U_1 h_2}{U_2 - U_1} \tag{13}$$

where the subscripts designate the two height/wind pairs. Assuming $\Delta h_1 = \Delta h_2 = \Delta h$, the RMS error in h_o is

$$\Delta h_o = \frac{\Delta h (U_2^2 + U_1^2)^{1/2}}{U_2 - U_1} \tag{14}$$

α is determined by the slope of the line connecting the two points:

$$\alpha = \frac{h_2 - h_1}{U_2 - U_1} \tag{15}$$

so that

$$\Delta\alpha = \frac{\sqrt{2}\,\Delta h}{U_2 - U_1} \tag{16}$$

The other quantity in Equation 10, h_m is subject to errors based on the type of measurement, the method of derivation, and also the time of day. h_m can be measured by incoherent scatter radar. This has the potential of being the most accurate, although the accuracy depends on various radar parameters such as the pulse length, signal-to-noise ratio, the conversion between scattered power and electron density, and the method of interpolation used to determine the maximum in the ionization density.

The majority of the h_{max} measurements on the global scale are made by ionosondes. The newer ionosondes are able to record the ionogram digitally and convert it to an electron density profile using a true-height analysis such as POLAN [Titheridge et al., 1985]. The majority of the ionosondes are still scaled by hand. The level of effort required to analyze an ionogram using a true-height analysis for hand-scaled ionograms is so great as to be impractical except for special campaigns.

The best alternative to a true-height analysis is to use an empirical formula, such as Equation 11, that determines the layer height from *M(3000)F2*. There have been few studies published of the accuracy of the *M(3000)F2* technique. Dudeney [1983] suggests that the accuracy in the derived h_m is about 5%.

Figure 3 shows the derived meridional wind at Tokyo on 18 September 1984, during the Equinox Transition Study. The error bars in Figure 3 are the statistical error bars computed using Equation 12. The error bars are small compared with the diurnal variation of the derived winds, and are of comparable size to the hourly variability on this particular day. The winds are derived from data taken on a magnetically quiet day, and would normally be much more variable.

The greater sources of uncertainty are not statistical, however, but are systematic and derive from the lack of knowledge of many of the physical processes involved. The model used, while it represents a detailed photochemical description of the thermosphere and ionosphere, is representative of average conditions. Recently Burnside et al. [1987] has shown evidence that the 0-0^+ collision cross section should be increased by a factor of approximately 1.7 over the previously used value. An increase in collision cross section in the FLIP model has the effect of significantly reducing the magnitude of the derived winds. Unless otherwise stated, the model results illustrated in this paper were derived using the larger collision cross section.

The magnitude of the east-west electric field also introduces a systematic error into the derived winds (Equation 9). There is currently no electric field model that has been shown to be accurate on a global scale, although some have successfully reproduced local measurements. This is, however, not a serious limitation if the results of the wind derivation is used as input to an ionospheric model, since the sum of the forces due to the meridional neutral wind and the electric field is the quantity needed by most models.

4.4. Dependence of h_0 and α on Latitude, A_p, and F10.7

The response of the height of the *F2* layer to changes in the meridional neutral wind speed depends strongly on the inclination of the geomagnetic field and of the local atmospheric conditions. As discussed above, the main determining factors of the layer height are the component of the neutral wind that is parallel to the magnetic field, diffusion of the ions along the field, and recombination.

Figure 4 shows the local time variation of α and h_0 at four different magnetic latitudes, as computed by the FLIP model. The calculations were done for magnetically

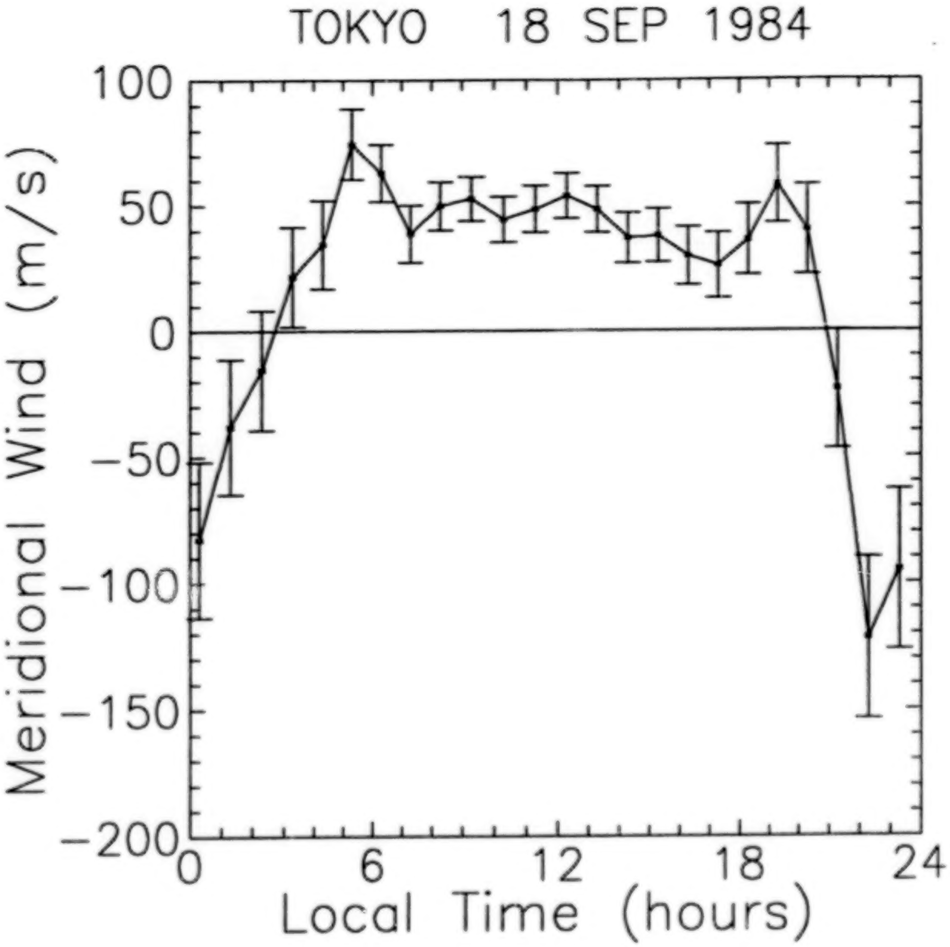


Figure 3. Meridional wind speed at Tokyo (36° N, 140° E) on 18 September 1984. Error bars were computed using Equation 12.

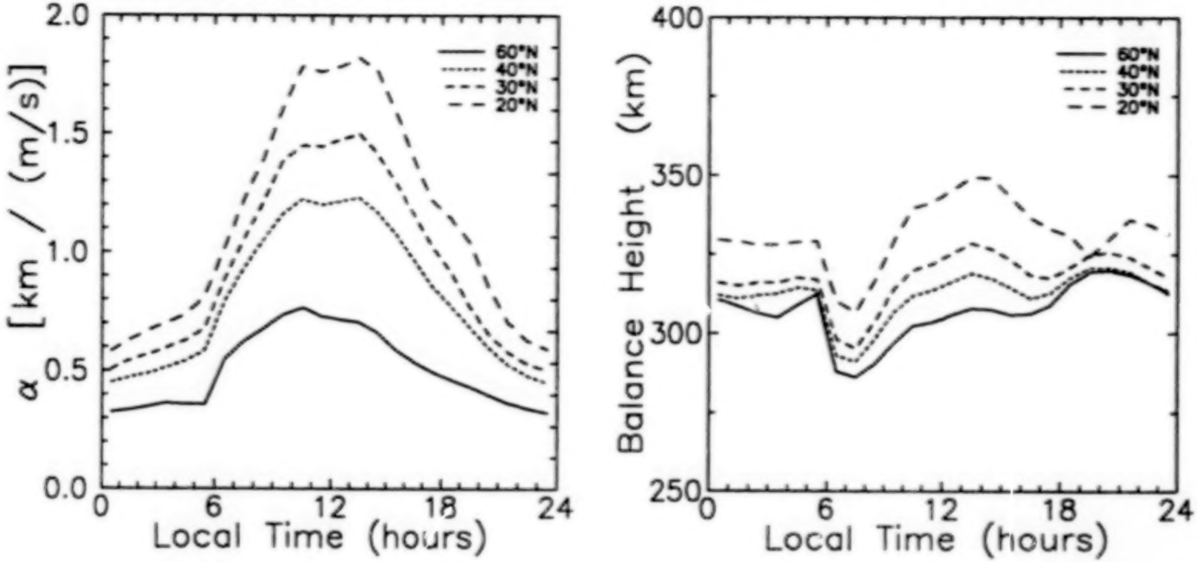


Figure 4. Local time variation of α and h_o at four geomagnetic latitudes.

quiet conditions ($A_p = 5$) and moderate solar flux (F10.7 = 130) at a magnetic longitude of 0°.

A simple expression for the effect of latitude on α can be derived by assuming that the effect of a neutral component parallel to the magnetic field has a constant effect on the position of the peak of the layer along the field line at a given local time. In other words, at any given time,

$$\alpha_{||} = \Delta h_{||}/\Delta U_{||} \cong \text{constant}, \tag{17}$$

where the subscripts are to indicate that only the dimension parallel to the magnetic field is to be considered. The distance along the field line (h_2-h_1 in Equation 15) can be related to vertical distance by

$$\Delta h_m = \Delta h_{||} \sin(I) \tag{18}$$

and the component of the horizontal neutral wind along the field line to the horizontal wind by

$$\Delta U = \Delta U_{||} / \cos(I) \tag{19}$$

so that the value of α that relates vertical layer height to horizontal wind changes can be written in terms of $\alpha_{||}$:

$$\alpha = \alpha_{||} \sin(I) \cos(I) \tag{20}$$

This relationship was used by Forbes et al. [1988] to derive meridional wind speed from ionosonde data in a study of the penetration of auroral effects to low latitude during a magnetic storm. They fit the form of $\alpha_{||}$ published by Miller et al. [1986] to a mathematical function and scaled it by sin(I)cos(I) to apply it to the latitudes of a meridional chain of ionosonde stations through Japan and the Soviet Union. Figure 5 shows $\alpha_{||}$ derived from the curves shown in Figure 4 using the dip angle from the tilted dipole magnetic field model used in the FLIP model. This seems to be a fair approximation at mid-latitudes, but departs from the calculated values at low latitudes.

Magnetic latitude is the primary geographic parameter in determining α and h_0. Even though the ionospheric density is affected by the neutral atmosphere, the effect of geographic latitude on the derivation of winds is much less important than the effect of magnetic latitude. Figure 6 shows the two parameters at four different longitudes, but all at a magnetic latitude of 40° North. Even though the geographic latitude varies from 29° to 51° North, there is very little difference in α or h_0.

Variation of α and h_0 with respect to magnetic activity is shown in Figure 7. The calculations for Figure 7 and 8 are for 40° N, 0° E, magnetic coordinates. The variation is primarily in the balance height, although there is a small change in α. The insensitivity of these parameters to magnetic activity has important consequences in the study of storm effects, since they occur quickly and it is difficult to model the ionosphere with adequate time resolution.

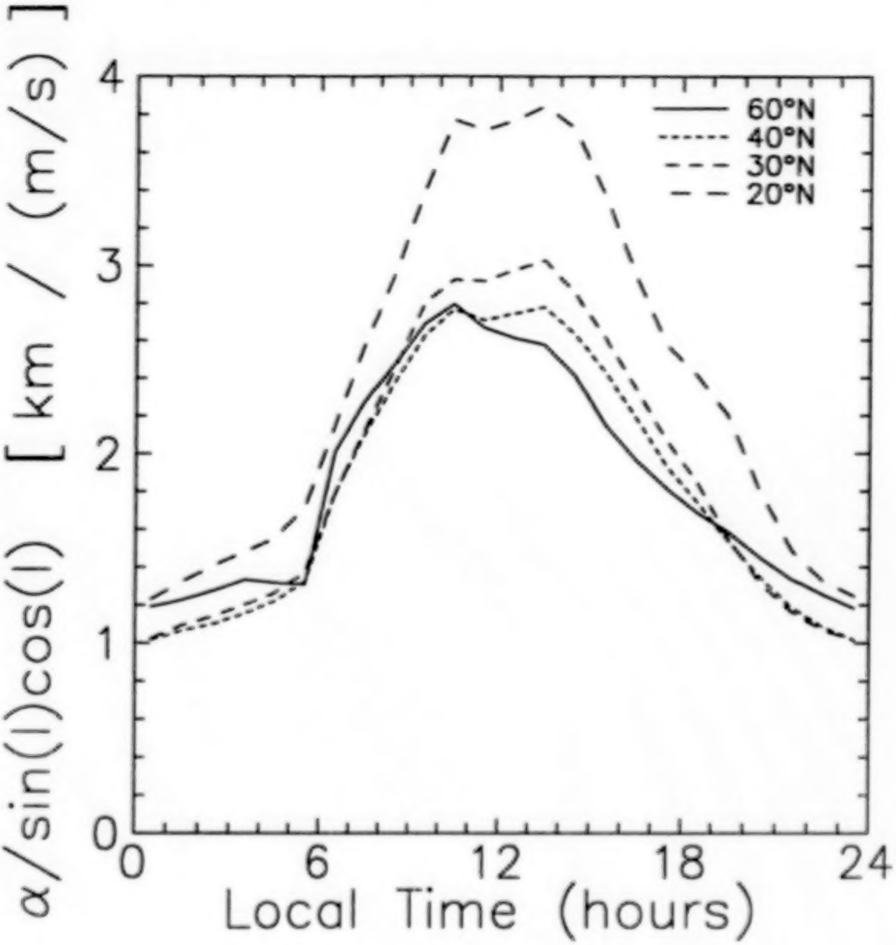


Figure 5. $\alpha_{\parallel}$ computed from values of α shown in Figure 4, and using the dip angle I from the tilted dipole magnetic field used in the FLIP model.

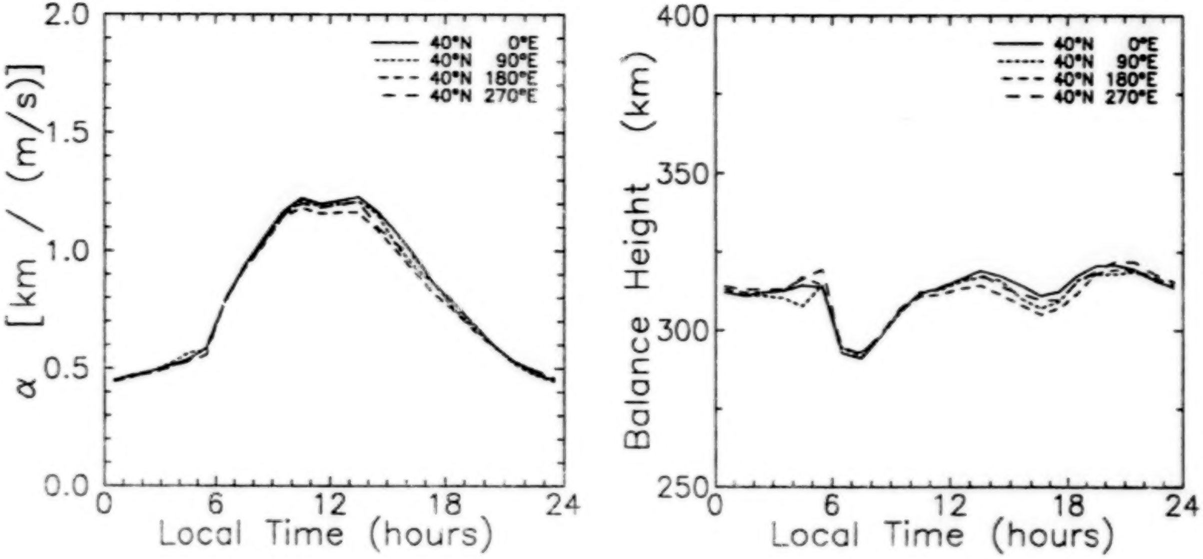


Figure 6. Local time variation of α and h_o at four geomagnetic longitudes. Magnetic latitude is 40° N.

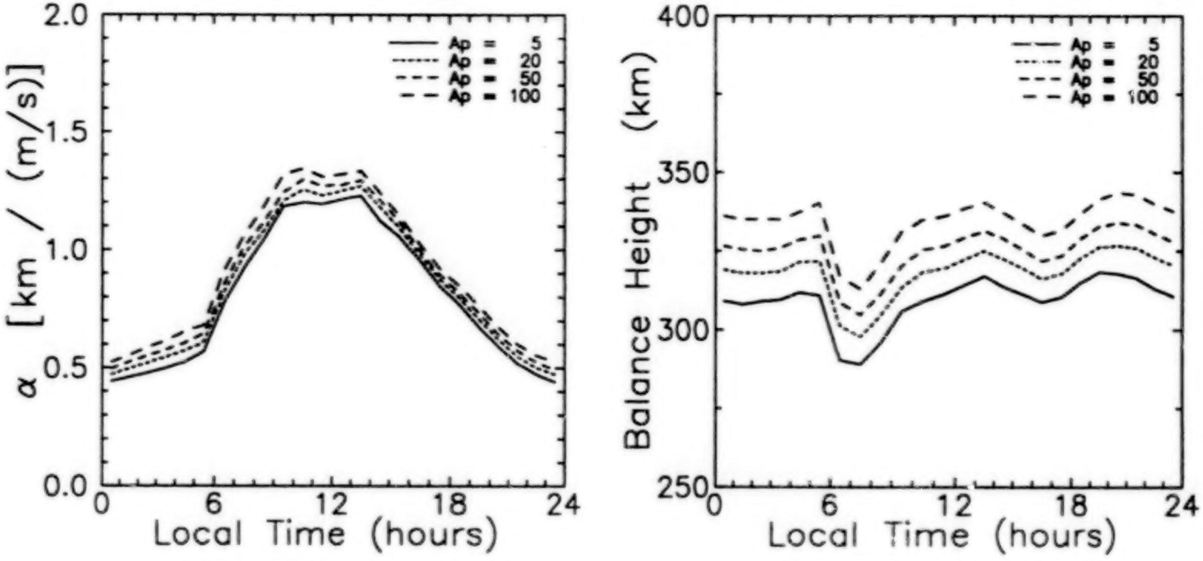


Figure 7. Local time variation of α and h_o with respect to magnetic activity.

A larger effect is found in the variation of α and h_o with changes in solar flux. Figure 8 shows α and h_o at low magnetic activity (A_p = 5) and at F10.7 values representative of conditions from solar minimum (F10.7 = 70) to solar maximum (F10.7 = 190). There are major differences in both α and h_o over this range of F10.7 values, mainly the result of the increased production rate of ionospheric O^+ and the larger scale height of the warmer neutral atmosphere at solar maximum. Although α and h_o are both strongly dependent on F10.7, it is such a slowly-varying function that it does not introduce a large uncertainty into the technique for deriving winds. While it will change significantly from month to month, there is very little day-to-day variation, even during magnetic storms.

5. CURRENT RESULTS

5.1 Comparisons With Other Wind Measurements

Several comparisons have been made of winds derived from h_{max} with meridional winds derived from other methods. These have been primarily with incoherent scatter radar, although Miller et al. [1986] also compared results with winds measured by Fabry-Perot interferometry. In each of these comparisons the wind speeds derived from h_{max} are consistent with the results of the other technique, both in the magnitude and in the pattern of variability of the winds.

Plots of the comparisons from Miller et al. [1986] are shown in Figure 9. The winds were measured at Arecibo using Fabry-Perot interferometry (F.P.I.) and derived from incoherent scatter ion drift measurements (v_B), and published by Burnside et al. [1983]. The curves labeled "Model" are derived from h_{max} without the 1.7 factor in the collision cross section [Burnside et al, 1987].

Miller et al. [1987] made a detailed comparison with the Millstone Hill incoherent scatter radar. This study had two objectives: First, to compare the difference between the results with the statistical uncertainty of the two derivations of wind speed. Second, to include measured electric fields in the calculation of winds from h_{max} to see if this would account for the difference.

The comparison was made for the Global Thermospheric Mapping Study (GTMS) campaigns of 26-18 June 1984 and 15-17 January 1985. Meridional neutral winds were derived from ion drift measurements for both campaigns. East-west electric fields were measured only during the June campaign. The factor of 1.7 was not used in the collision cross section in either wind derivation of this study.

Figure 10 shows the winds derived from h_{max} and from ion drifts for the two GTMS campaigns. The dashed curves are from incoherent scatter radar measurements of ion drift; the solid curves are derived from measurements of h_{max}. In this case, radar measurements of h_{max} were used. The results of the two derivations are quite similar, both in the magnitude of the wind and in the correlation of the features. An example is the abatement in the southward wind that occurs near sunset in January.

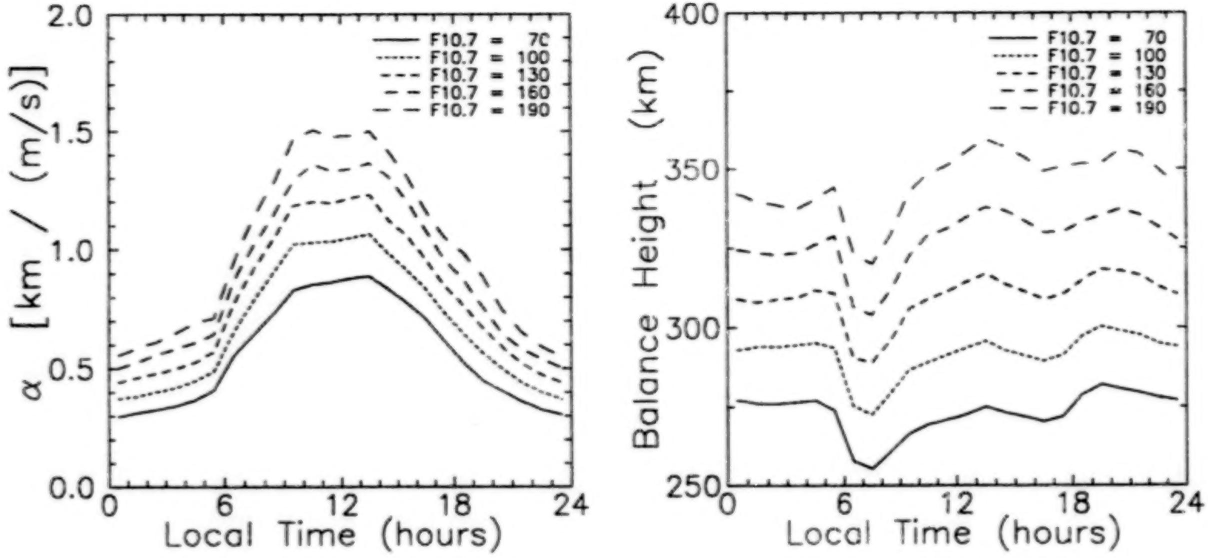


Figure 8. Local time variation of α and h_o with respect to changes in solar flux. Range of F10.7 indices spans the expected difference from solar minimum to solar maximum.

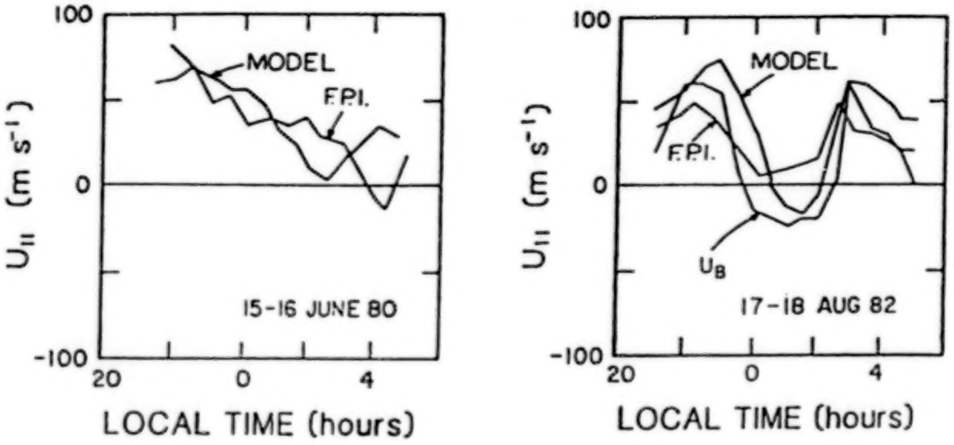


Figure 9. Comparisons of winds from h_{max} with winds measured by Fabry-Perot interferometry (F.P.I.) and by incoherent scatter ion drift measurements (U_B) for Arecibo [Miller, et al., 1986]

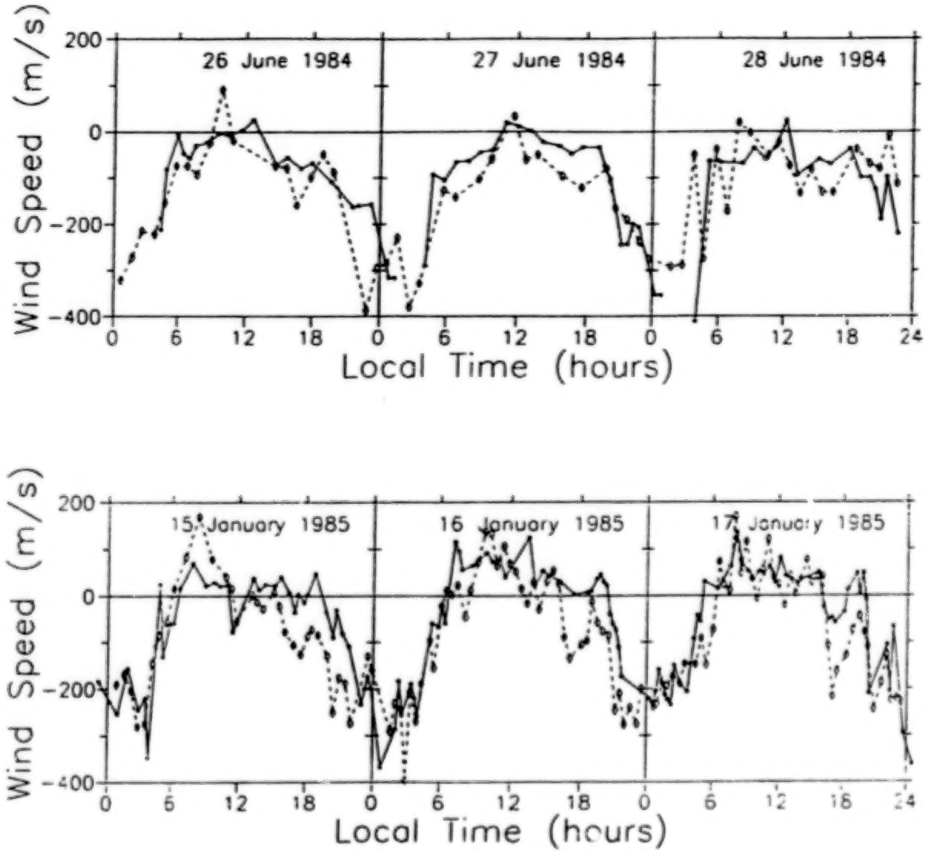


Figure 10. Comparisons of winds from h_{max} with winds derived from ion drift velocity measurements for 26-28 June 1984 and 15-17 January 1985. Dashed curves are from ion drift; solid curves are from h_{max}.

An attempt was made to determine the sources and magnitudes of the errors involved in these two derivations to see if the differences in the results are significant. The difference between the wind from h_{max} and the radar result was the greatest on 27 June 1984. The statistical error bars, shown in Figure 11, assume a 15-km uncertainty in the radar measurement of h_{max}. Figure 11 also shows the winds measured by the radar with their statistical error bars. With the exception of the afternoon near sunset the results are not significantly different. There are common sources of error inherent in both of these calculations. They both rely on model calculations, and are subject to the uncertainties inherent in models. An example is the uncertainty in the value of the O^+-O collision cross section.

Both the MSIS atmospheric model and the FLIP ionospheric model have been tested against many data sets, and have been shown to be reliable. However, a model, of necessity, must reflect average conditions. When daily departures of the models from actual conditions are considered, as well as uncertainties in the parameters used in the model calculations of both diffusion velocity and layer height, a more realistic error bar might be as large as 50 to 75 m/s. If 75 m/s error bars were used in Figure 11, the two wind curves would be within the error range of one another.

The east-west electric field was derived from ion velocity measurements made at Millstone Hill during the June campaign. The measured field, shown in Figure 12, was small throughout most of the time interval. The largest electric fields, reaching 4 mV/m in a westward direction, were measured on June 27.

Using Equation 9, we can determine the relative magnitude of the error introduced by neglecting the electric field. Figure 13 illustrated the results of including this measured field in the wind calculation. In Figure 13, the dashed curve is the wind from the incoherent scatter measurement. The solid curve was derived from h_{max}, and the dotted curve is the wind from h_{max} corrected by the electric field measurement, as described by Equation 9. On June 27, when the strongest electric fields were measured, the difference between the curves in the morning is accounted for by the 4 mV/m field. However, the measured electric field is not enough to reconcile the two wind curves in the afternoon.

Winds derived from ionosonde measurements cannot be independently verified. However, ionosondes near incoherent scatter radars should give similar results to the radars. An example is shown in Figure 14, where the meridional wind derived from ion drift measurements by the Millstone Hill radar are compared with the winds from h_{max} measurements by the Ottawa ionosonde for 26 June 1984. The two stations are separated by about 500 km in distance. Millstone Hill is at 42.5° N, 71.5° W, and Ottawa is at 45.1° N, 76.1° W.

The three-parameter formulation developed by Dudeney [1983] (Equation 11) was used to find the ionosonde values of h_{max}. There is agreement in both the wind speed and in the times of occurrence of the major variations at the two locations. The major feature is the abatement in the southward wind that occurs near sunset. This abatement was also seen in the wind values for the January GTMS period at Millstone Hill.

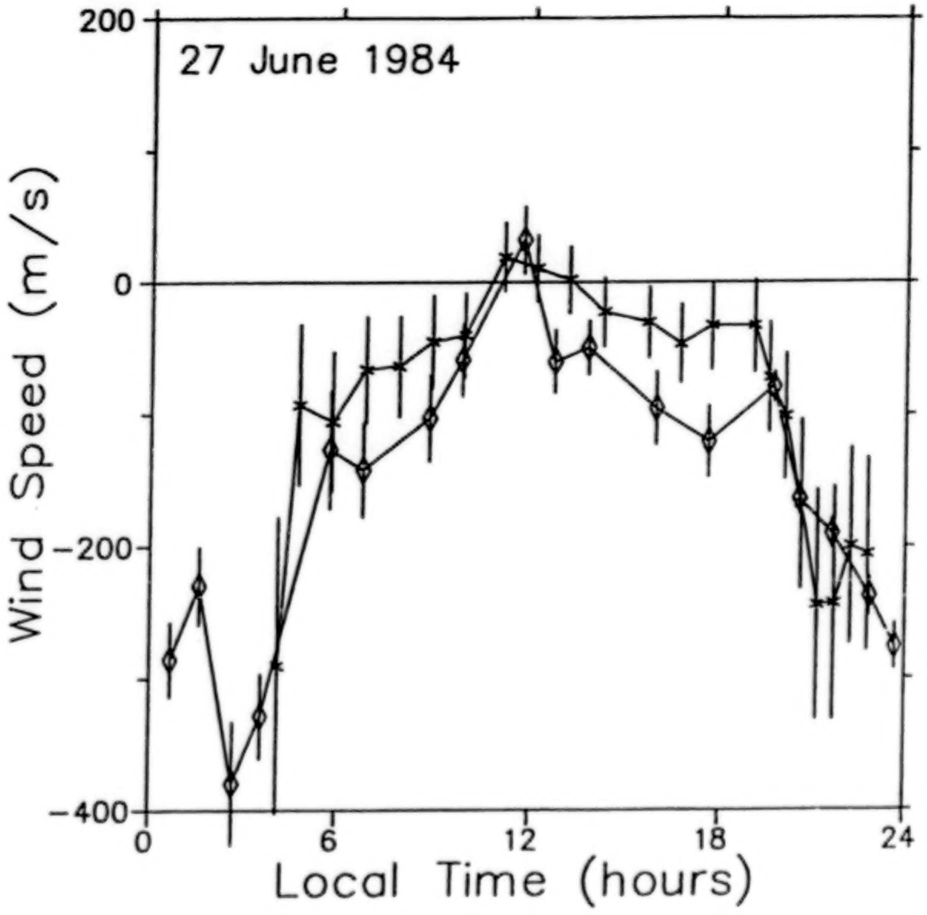


Figure 11. Meridional neutral wind speed at Millstone Hill, 27 June 1984, showing statistical uncertainty of the two derivations of wind speed. Diamonds show speed derived from ion drift measurements. Crosses are winds from h_{max}.

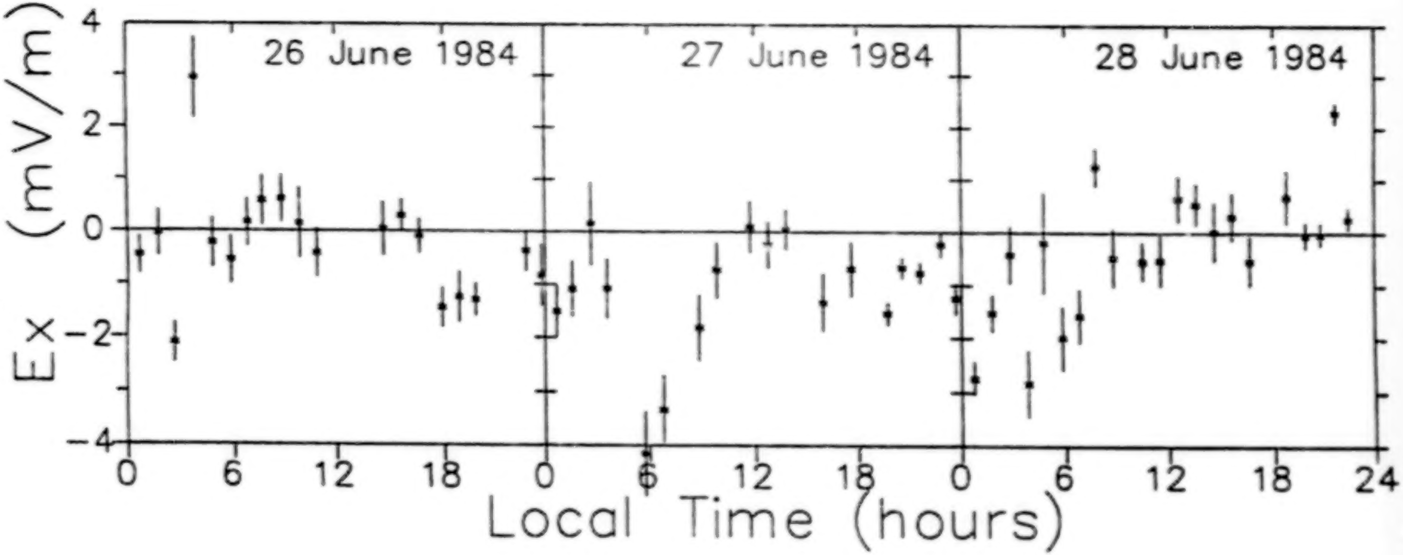


Figure 12. Eastward electric field measured by the Millstone Hill Radar during the June, 1984, GTMS campaign.

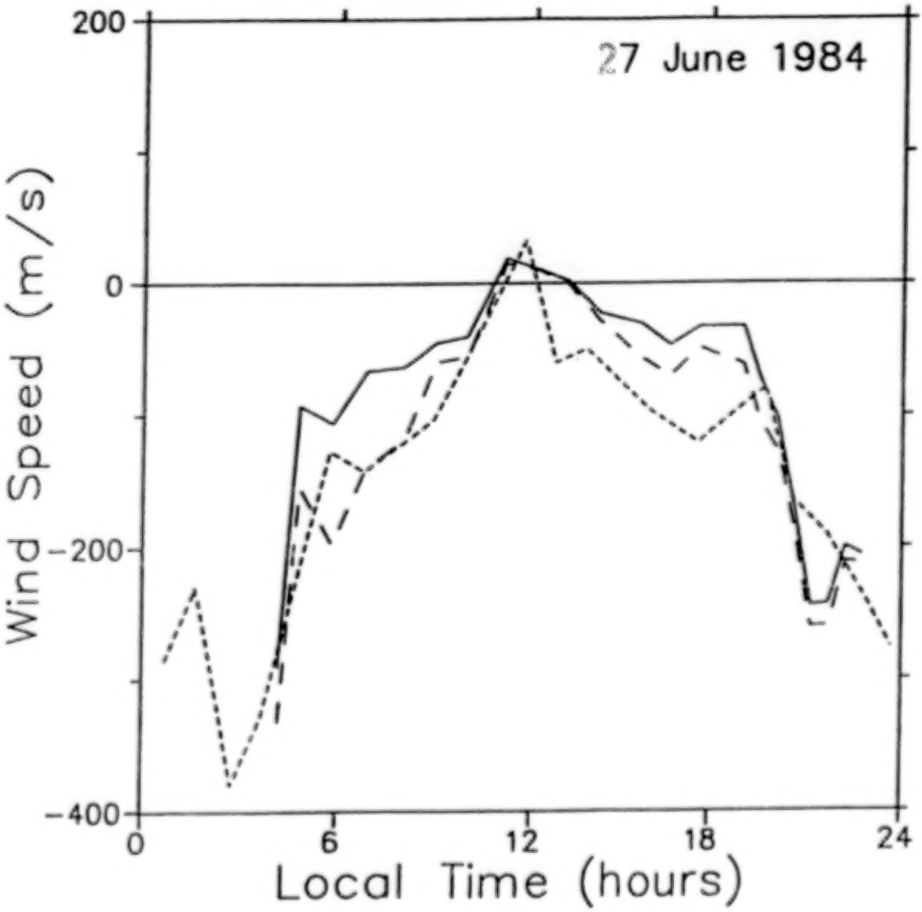


Figure 13. Wind speed at Millstone Hill corrected for the effects of electric fields. The solid line shows wind speed from h_{max}; long dashes are wind from h_{max} with electric fields included; short dashes show wind speed from incoherent scatter ion drift measurements.

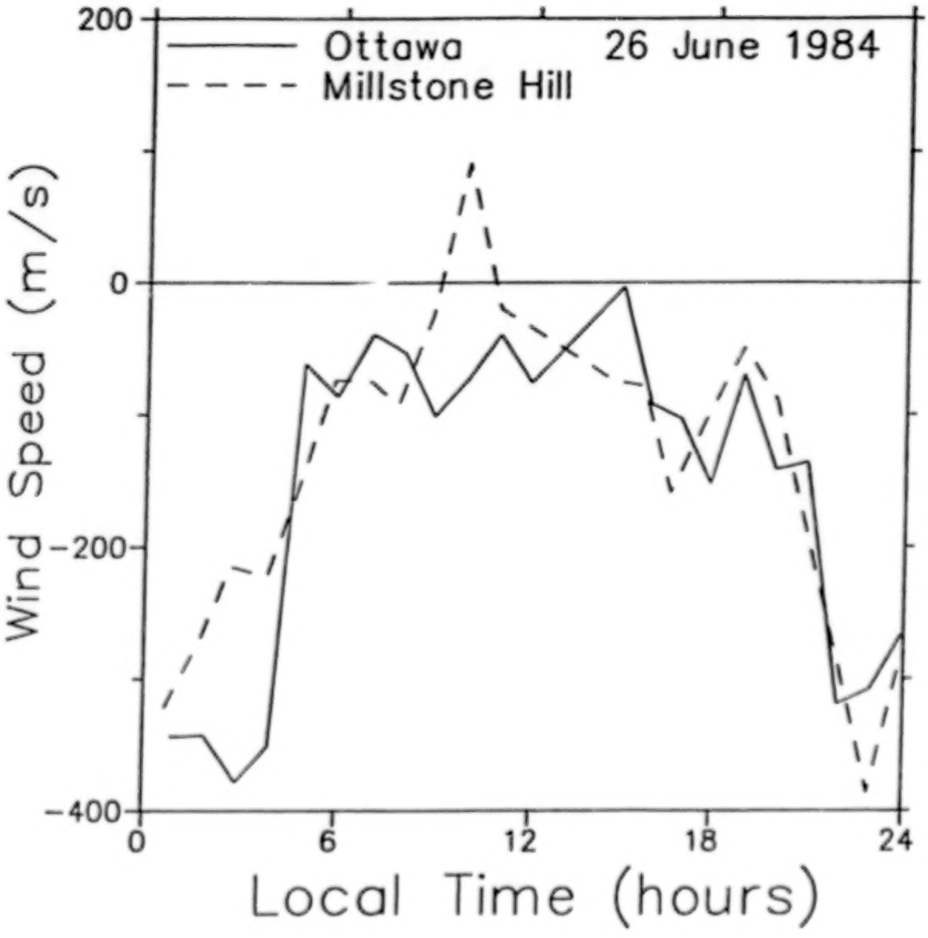


Figure 14. Meridional wind derived from ion drift measurements by the Millstone Hill Radar compared with the winds from h_{max} measurements by the Ottawa ionosonde for 26 June 1984. The two stations are separated by about 500 km in distance. Millstone Hill is at 42.5° N, 71.5° W, and Ottawa is at 45.1° N, 76.1° W.

A comparison of winds derived from h_{max} with results of the NCAR TGCM [Crowley, et al., 1989] is shown in Figure 15. This example is from the Equinox Transition Study of September 1984. Magnetic conditions were quiet on the first day, but the second day includes an equatorward surge at the onset of a magnetic storm. This was a "blind" comparison; no attempt was made to adjust either model to make it agree with the other.

5.2 Comparison With the Servo Model

In a recent study, Buonsanto et al. [1989] derived winds from an extensive Millstone Hill Radar data set using both the servo model and the FLIP model. The radar was being operated in a mode that used full north-to-south elevation scans for about seven days. The meridional neutral winds were derived from h_{max} from 30° to 55° latitude for seven days using both techniques. Although the derived winds are similar, servo-model winds are consistently more negative (stronger southward) than the winds derived using the FLIP model. The ratio of the two models is about 1.2 at night, and drops to near 0.5 in the daytime. Some of the difference may be due to the difference between the daytime and nighttime servo equations (see Equations 4 and 5), although a gradual transition is made between them. Another source of difference between the two techniques is the calculation of balance height. Also, in this case, the servo model used electron and ion temperatures that were derived from incoherent scatter radar measurements, while the FLIP model calculated the temperatures.

It would be convenient to be able to use a servo-type model to derive the wind values. However, the difference between the two techniques is significant. At this writing, an examination of the assumptions made in the two techniques and a comparison of them under a variety of conditions is being done to reconcile the results of the two techniques.

5.3 Support of Measurement Campaigns

Efforts to study ionosphere-thermosphere coupling are being coordinated through CEDAR on a national level, and through WITS and an international level. Both programs are interested in global wind values during campaigns. To date, the GTMS campaigns of June, 1984, and January, 1985; the ETS campaign of September, 1984; each of the GITCAD campaigns, and all but the first SUNDIAL campaign are supported by global-scale derivation of meridional wind speeds. Since the data required for the derivation of h_{max} are collected and archived routinely, it is possible to derive global scale meridional winds for any campaign.

5.4 Investigation of Solar Cycle Effects.

Breninger [1989], Miller et al. [1989], and Buonsanto [1990] have reported long-term trends in the meridional winds based on measurements of h_{max}. Breninger and Miller et al. examined data from several ionosondes over the last solar cycle and derived $hmF2$ values for the ten quietest days of January over an eleven-year period. Buonsanto used data from the ionosondes at Boulder and Wallops Island, and looked at diurnal, seasonal, and solar cycle variations. Each study reported a significant variation in the diurnal

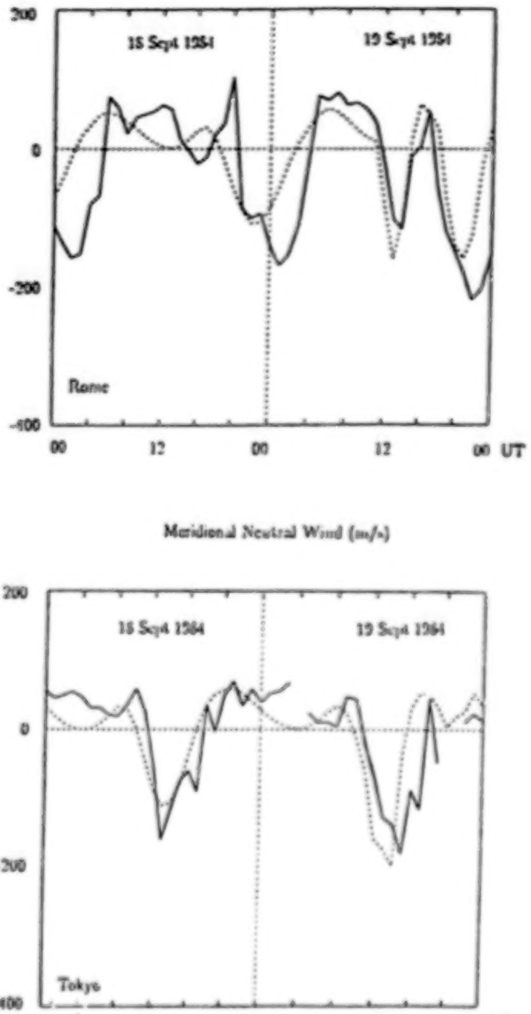


Figure 15. Meridional neutral winds measured (solid) and predicted (broken) for Rome and Tokyo. Negative values indicate equatorward winds; units are m s^{-1} [Crowley et al., 1989].

amplitude of the wind variation at the northern (winter) hemisphere stations and attributed it to the increased ion drag on the neutral atmosphere at solar maximum. There is apparently a weaker variation, or at least a more confused situation, at low latitudes and in the southern (summer) hemisphere. In general, winds were strongest during solar minimum, but did not vary in the same pattern as the 10.7 cm flux or the sunspot number, as shown in Figure 16.

5.5 Development of a Global Wind Model.

As the reliability of the derivation of meridional neutral winds from ionosondes becomes established for quiet times, it will be possible to develop an empirical neutral wind model based entirely on ionosonde measurements. The size of the data base is already large enough to provide adequate diurnal and solar cycle coverage, although allowance must be made for the lack of coverage in some parts of the globe.

The possibility of a global wind model based on ionosonde measurements is illustrated by a study by Miller et al. [1990] to compare the winds derived from median ionosonde data for a 15-day period with winds derived from the International Reference Ionosphere (IRI) values of *hmF2*. Figure 17 shows agreement between ionosonde values and IRI values of the meridional wind speed at a set of ionosonde stations at approximately constant longitude. The stars show wind speed derived from 15-day median ionosonde measurements of h_{max}. The solid line is the wind speed derived from IRI values of *hmF2*. The dashed line that is included in Figure 17 is the wind speed from the HWM87 empirical wind model [Hedin et al, 1988]. The HWM87 model is based on a vector spherical harmonic fit to satellite wind measurements.

Comparing winds derived from ionosonde measurements with winds derived from the IRI is equivalent to comparing the IRI predictions of *hmF2* with ionosonde measurement. However, it suggests the possibility of developing a global wind model based on the IRI. The reliability of such an empirical meridional wind model would be critically dependent on the reliability of the IRI values of *hmF2*. The IRI uses CCIR maps of ionosonde data to generate *hmF2* values. These maps are, of course, most reliable in areas where the ionosonde measurements are most extensive, such as Europe and the Japanese-Australian meridian. To verify the accuracy of the IRI would be a long-term project, possibly using *hmF2* and neutral wind measurements from satellites.

Acknowledgements: This work was supported by NSF Grant ATM-87-15367 to Utah State University and by NSF Grant ATM-87-16036 to the University of Alabama in Huntsville.

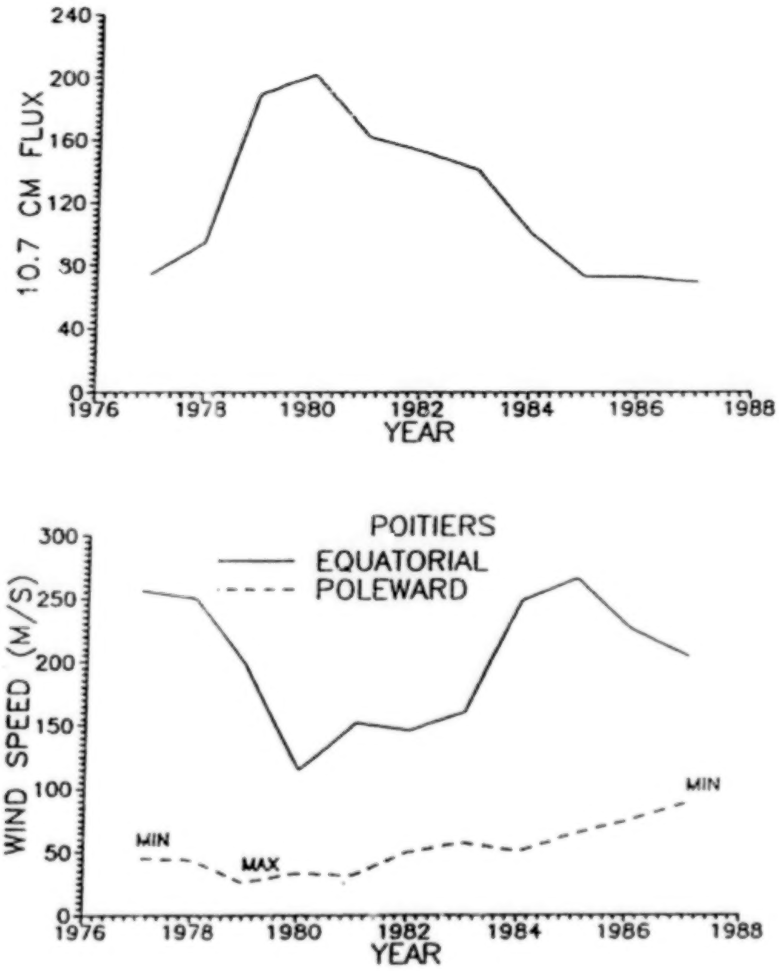


Figure 16. Variation of the equatorward (nighttime) and poleward (daytime) meridional winds for one solar cycle relative to the 10.7 cm solar radio flux. Units of flux are 10^{-22} W m^{-2} Hz^{-1} [Breninger, 1989].

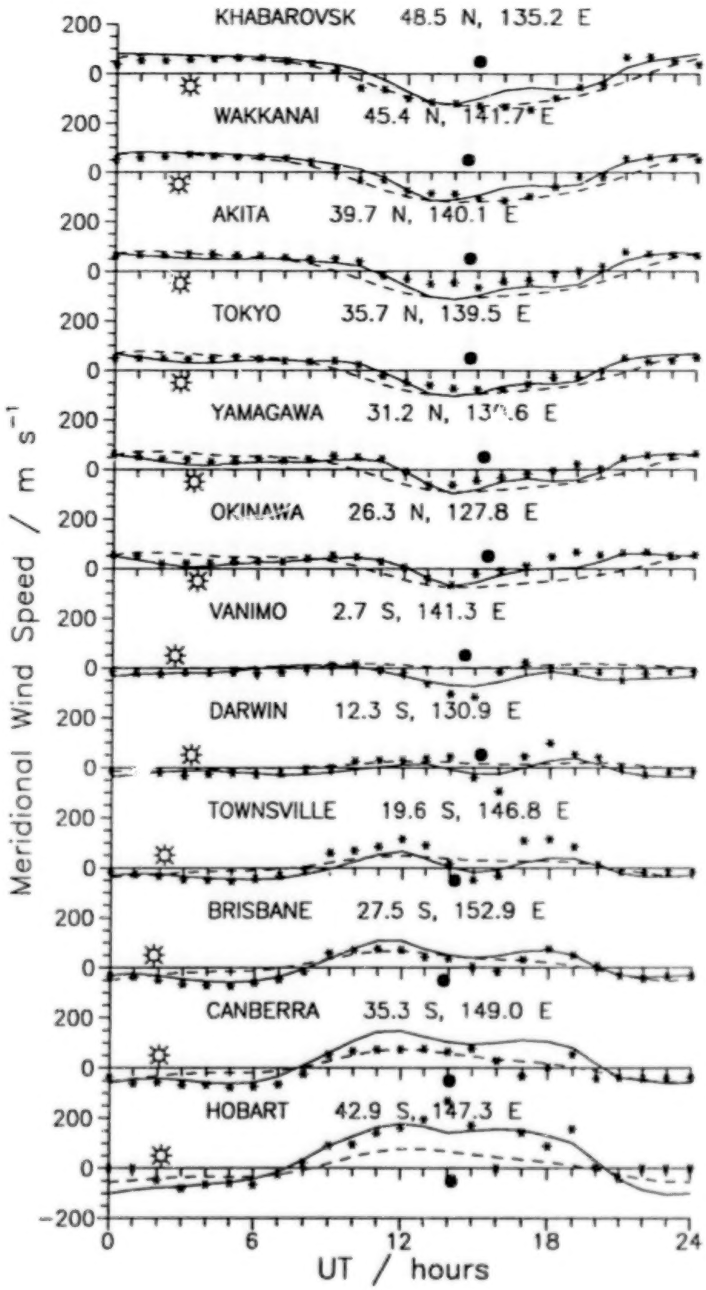


Figure 17. Meridional neutral wind speeds (m s^{-1}) at a meridional chain of ionosonde stations. Positive winds are northward. Stars are 15-day medians of ionosonde winds; solid line, IRI winds; and dashed line, HWM87 winds. Open and shaded circles mark local noon and midnight. [Miller, et al, 1990].

REFERENCES

Bradley, P. A., and J. R. Dudeney, A simple model of the vertical distribution of electron concentration in the ionosphere, J. Atmos. Terr. Phys., 35, 2131-2146, 1973.

Breninger, R. L., Solar cycle trend study of the mid-latitude quiet time meridional neutral winds at winter solstice conditions, M.S. Theses, Utah State University, 1989.

Buonsanto, M. J., J. E. Salah, K. L. Miller, W. L. Oliver, R.G. Burnside, and P. G. Richards, Observations of neutral circulation at mid-latitudes during the equinox transition study, J. Geophys. Res., 94, 16,987-16,997, 1989.

Buonsanto, M. J., Observed and calculated *F2* peak heights and derived meridional winds at mid-latitudes over a full solar cycle, J. Atmos. Terr. Phys., in press, 1990.

Burnside, R. G., R. A. Behnke, and J. C. G. Walker, Meridional neutral winds in the thermosphere at Arecibo: Simultaneous incoherent scatter and airglow observations, J. Geophys. Res., 88, 3181-3189, 1983.

Burnside, R. G., C. A. Tepley, and V. B. Wickwar, The O^+-O collision cross-section: Can it be inferred from aeronomical measurements?, Annales Geophysicae, 5A, 343-350, 1987.

Chandler, M. O., R. A. Behnke, A. F. Nagy, E. G. Fontheim, P. G. Richards and D. G. Torr, Comparison of measured and calculated low-latitude ionospheric properties, J. Geophys. Res., 88, 9187-9196, 1983.

Crowley, G., B. A. Emery, R. G. Roble,, H. C. Carlson, Jr., J. E. Salah, V. B. Wickwar, K. L. Miller, W. L. Oliver, R. H. Burnside, and F. A. Marcos, Thermospheric dynamics during September 18-19, 1984. 1. Model simulations, J. Geophys. Res., 94, 16,925-16,944, December 1989.

Dudeney, J. R., The accuracy of simple methods for determining the height of the maximum electron concentration of the *F2*-layer from scaled ionospheric characteristics, J. Atmos. Terr. Phys., 45, 629-640, 1983.

Forbes, J. M., M. Codrescu and T.J. Hall, On the utilization of ionosonde data to analyze the latitudinal penetration of ionospheric storm effects, Geophys. Res. Lett., 15, 249-252, 1988.

Hanson, W. B., and T, N. Patterson, The maintenance of the night-time *F*-layer, Planet. Space Sci., 12, 979-997, 1964.

Hedin, A. E., N. W. Spencer, and T. L. Kileen, Empirical global model of upper thermosphere winds based on Atmosphere and Dynamics Explorer satellite data, J. Geophys. Res., 93, 9959-9978, 1988.

Miller, K. L., D. G. Torr, and P. G. Richards, Meridional winds in the thermosphere derived from measurement of F2 layer height, J. Geophys. Res., 91, 4531-4535, 1986.

Miller, K. L., J. E. Salah, and D. G. Torr, The effect of electric fields on measurements of meridional neutral winds in the thermosphere, Annales Geophysicae, 5A, 337-342, 1987.

Miller, K. L., R. L. Breninger, P. G. Richards, and D. G. Torr, The solar cycle variation of meridional neutral winds in the thermosphere, EOS, 70, 1243, 1989.

Miller, K. L., A. E. Hedin, P. J. Wilkinson, D. G. Torr, and P. G. Richards, Neutral winds derived from IRI parameters and from the HWM87 wind model for the SUNDIAL campaign of September, 1986, Adv. Space Res. 10, (8)99-(8)102, 1990.

Peddie, N. W., International Geomagnetic Reference Field: The third generation, J. Geomag. and Geoelec., 34, 309-326, 1982.

Richards P. G. and D. G. Torr, Seasonal, diurnal, and solar cycle variations of the limiting H^+ flux in the earth's topside ionosphere, J. Geophys. Res., 90, 5261-5268, 1985.

Richards, P. G. and D. G. Torr, Ratio of photoelectron to EUV ionization rates for aeronomic studies, J. Geophys. Res., 93 4060, 1988.

Rishbeth, H., and D. W. Barron, Equilibrium electron distributions in the ionospheric F2-layer, J. Atmos. Terr. Phys., 18, 234-252, 1960.

Rishbeth, H., *F2*-layer rates at sunspot minimum, J. Atmos. Terr. Phys., 28, 911-918, 1966.

Rishbeth, H., The effects of winds on the ionospheric *F2*-peak, J. Atmos. Terr. Phys., 29, 225-238, 1967.

Rishbeth, H., Thermospheric winds and the *F*-region: A review, J. Atmos. Terr. Phys., 34, 1-47, 1972.

Rishbeth, H., S. Ganguly, and J. C. G. Walker, Field-aligned and field-perpendicular velocities in the ionospheric *F2*-layer, J. Atmos. Terr. Phys., 40, 767-784, 1978.

Shimazaki, (1955), J. Radio Research Labs, Japan, 2, 85, 1955.

Sojka, J. J., Global scale, physical models of the *F* region ionosphere, Rev. Geophys, 27, 371-403, 1989.

Titheridge, J. E., Ionogram analysis with the generalized program POLAN, Report UAG-93, World Data Center A for Solar-Terrestrial Physics, NOAA, E/GC2, Boulder, CO, 80303, 1985.

Torr, D. G., P. G. Richards, and M. R. Torr, Photochemistry of the ionosphere, WITS Handbook, 1, 1-38, 1988.

Young, E. R., P. G. Richards, and D. G. Torr, A flux preserving method of coupling first and second order equations to simulate the flow of plasma between the photosphere and the ionosphere, J. Comp. Phys., 38, 141-155, 1980a.

Young E. R., D. G. Torr, P. Richards and A. F. Nagy, A computer simulation of the midlatitude plasmasphere and ionosphere, Planet. Space Sci., 28, 881, 1980b.

☆ U.S. GOVERNMENT PRINTING OFFICE 1990—775-386 25109

WITS HANDBOOK VOLUME 2
EXPERIMENTAL TECHNIQUES

Contents

Chapter Page

1. Use of ground-based magnetometers to infer the dynamics of magnetosphere-ionosphere coupling and the solar-terrestrial interaction, *G. Rostoker* 1

2. "In situ" measurement techniques for ionospheric-thermospheric investigations, *E. P. Szuszczewicz* 44

3. Prospects of topside sounding, *S. A. Pulinets* 99

4. Scintillation technique for probing ionospheric irregularities, *Sa. Basu and Su. Basu* 128

5. Modern total electron content measurement techniques, *J. A. Klobuchar* 137

6. Rayleigh and sodium lidar techniques for measuring middle atmosphere density, temperature and wind perturbations and their spectra, *C. S. Gardner, D. C. Senft, T. J. Beatty, R. E. Bills, and C. A. Hostetler* 148

7. Upper atmospheric wind and temperature measurements using imaging Fabry-Perot interferometers, *D Rees , I. McWhirter, A. Aruliah, and S. Batten* 188

8. Spectrometric measurement techniques of weak emissions in twilight airglow, *C. A. Tepley* 224

9. Optical and radar techniques applied to chemical research "active experiments" in the ionosphere/thermosphere system, *M. Mendillo, J. Baumgardner, and P J. Sultan* 250

10. Modification of the ionosphere as a method of its diagnostics, *L. M. Erukhimov and N. A. Mityakov* 267

11. Coherent radar techniques, *E. Nielsen* 285

12. Remote sensing of the high latitude ionosphere with coherent scatter HF radar: A review of the technique and recent work, *J. M. Ruohoniemi, R. A. Greenwald, and K. B. Baker* 319

13. The digisonde 256 ionospheric sounder, *B. W. Reinisch, K. Bibl, D. F. Kitrosser, G. S. Sales, J. S. Tang, Z.-M. Zhang, T. W. Bullett and J. A. Ralls* 350

14. The HF Doppler technique for monitoring transient ionospheric disturbances, *T. B. Jones* 383

15. Theoretical and simulation methods applied to high latitude, F region turbulence, *J. D. Huba* 399

16. The derivation of meridional neutral winds in the thermosphere from F2-layer height, *K. L. Miller, P. G. Richards, and D. G. Torr* 439

472